Étude de l'exposition de la population française au champ magnétique 50 Hertz

Mfoihaya Bedja

N° D'ORDRE : 9819

THÈSE DE DOCTORAT

SPECIALITE : PHYSIQUE

École Doctorale « Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIX XI

par

MFOIHAYA Bédja

ÉTUDE DE L'EXPOSITION DE LA POPULATION FRANÇAISE AU CHAMP MAGNÉTIQUE 50 HERTZ

Soutenue le 10 mai 2010 devant les membres du jury :

M. BENIDIR Messaoud	L2S, Gif-Sur-Yvette	Examinateur
M. FLEURY Gilles	Supélec, Gif Sur Yvette	Directeur de thèse
M. LILIEN Jean-Louis	Université de Liège, Belgique	Examinateur
M. LUTZ Pierre	CEA, Saclay	Rapporteur
Melle. MAGNE Isabelle	EDF R&D, Moret-Sur-Loing	Examinatrice
M. VEYRET Bernard	Laboratoire IMS CNRS/EPHE, Pessac	Rapporteur

EXPOSITION DE LA POPULATION FRANÇAISE AU CHAMP MAGNÉTIQUE 50 Hz

Bédja MFOIHAYA

28 juin 2010

Table des matières

Table des figures

Liste des tableaux

Remerciements

Comme le veut la tradition, je vais tenter de satisfaire au difficile exercice de la page des remerciements, peut-être la tâche la plus ardue de ces années de thèse. Non qu'exprimer ma gratitude envers les personnes en qui j'ai trouvé un soutien soit contre ma nature, bien au contraire. La difficulté tient plutôt dans le fait de n'oublier personne. C'est pourquoi, je remercie par avance ceux dont le nom n'apparaît pas dans cette page et qui m'ont aidé d'une manière ou d'une autre.

La première personne que je tiens à remercier est Laurent LE BRUSQUET, mon encadrant à SUPELEC, qui a su me laisser la liberté nécessaire à l'accomplissement de mes travaux, tout en y gardant un œil critique et avisé. Il a toujours montré de l'intérêt pour mes travaux et répondu à mes sollicitations lorsque le besoin s'en faisait sentir. J'espère que cette thèse sera un remerciement suffisant au soutien et à la confiance sans cesse renouvelée dont il a fait preuve à mon égard. Je remercie aussi Isabelle MAGNE et Martine SOUQUES, mes encadrantes à EDF et au Service des Etudes Médicales d'EDF de leur soutien technique et de tout ce qu'elles ont fait pour la réalisation de cette thèse. Je remercie Gilles FLEURY, mon directeur de thèse de m'avoir donné l'opportunité de réaliser cette thèse dans le département des Signaux et Systèmes Electronique de SUPELEC dont il est le chef. Je remercie aussi toute l'équipe du département et plus particulièrement Arthur TENENHAUS pour sa disponibilité à répondre mes questions mais aussi pour ses blagues. "Elles étaient parfois utiles !".

Je remercie Bernard VEYRET et Pierre LUTZ de m'avoir fait l'honneur d'être les rapporteurs de cette thèse et du regard critique, juste et avisé qu'ils ont porté sur mes travaux. J'éprouve un profond respect pour leur travail et leur parcours, ainsi que pour leurs qualités humaines. Je profite aussi de remercier les membres du jury Messaoud BENIDIR et Jean-Louis LILIEN d'avoir accepté de lire ma thèse et d'être présent à ma soutenance en tant qu'examinateurs.

Je remercie Didier-Dacunha CASTELLE d'avoir été toujours disponible pour répondre à mes questions et de m'avoir accompagné pendant mes re-

cherches de sujet de thèse.

Mes remerciements s'adressent plus particulièrement à ma femme madame Adjouza ASSOUMANI de m'avoir soutenu moralement et d'avoir été compréhensive tout au long de ces trois années de recherches. Elle m'a toujours soutenu pendant les moments les plus difficiles. Je remercie aussi ma sœur Manzel BEDJA de m'avoir aidé financièrement à partir à Madagascar pour commencer mes études supérieures à l'université d'Antananarivo mais aussi d'avoir accepté de s'occuper de notre mère à mon absence. Je remercie mon oncle Baoua ABDOU de m'avoir aidé financièrement pendant mon cursus scolaire.

Je remercie tous les membres de l'association AJDOF et plus particulièrement Hassani MCHANGAMA de m'avoir aidé administrativement à venir poursuivre mes études universitaires en France.
Mes remerciements s'adressent aussi à ma belle famille (monsieur et madame Assoumani DJAE, monsieur et madame Ali SAID, monsieur et madame Hadji SAID, monsieur et madame Lahadji SABITI, monsieur et madame Ahmed ASSOUMANI, monsieur et madame Saïd ASSOUMANI) de tout ce qu'elle a fait pour la soutenance de cette thèse, madame et monsieur HERI de m'avoir hébergé gratuitement pendant 3 ans, mon cousin Abdallah MOHAMED (Plasse) pour ses soutiens moraux et financiers, monsieur et madame Saïd ABDOU (Congolais), monsieur et madame Abdou PAPA, Saïd Mohamed MCHANGAMA pour leur générosité.

Pour finir, je remercie mes amis Kola, Rachid METAHRI, madame et monsieur Mohamed DJAE et Youssouf Mohamed AHAMADA, des amis qui sont toujours à mes côtés pendant les moments les plus difficiles.

Je dédie cette mémoire à ma mère Maria ABDOU et à mon père Bédja TABIBOU qui ont eu le courage de m'inscrire à l'école alors qu'ils n'avaient pas les moyens financiers. Ils m'ont toujours encouragé à faire des études en me disant : "si tu ne réussis pas, c'est toute la famille qui échoue".

Résumé :

Les champs magnétiques (CM) alternatifs de fréquence 50 Hz liés à l'électricité sont suspectés depuis une trentaine d'années d'être responsables de pathologies, plus précisément de leucémies chez l'enfant. Les dernières expertises collectives (WHO 2007, SCENHIR 2009) ont conclu que la dernière grande interrogation en ce qui concerne les CM d'extrêmement basse fréquence (ELF) est l'association statistique observée dans plusieurs analyses conjointes entre l'augmentation du risque de leucémie de l'enfant et une exposition aux CM supérieure à 0,4 μT en valeur moyenne sur 24 heures, sans qu'il n'existe de relation causale.

La thèse a pour objectif de caractériser l'exposition de la population française aux champs magnétiques 50 Hz, et en particulier d'identifier les principaux facteurs favorisant l'exposition. Deux échantillons (1000 enfants et 1000 adultes) représentatifs de cette population ont été construits en utilisant la méthode de tirage aléatoire. Chaque personne a porté un EMDEX II mesurant et enregistrant les CM auxquels elle est exposée pendant une durée de 24 heures et a, au fur et à mesure, rempli un emploi du temps ainsi qu'un questionnaire contenant des informations personnelles et des informations relatives à son foyer. Lors de la récupération de l'appareil, l'enquêteur a enregistré les coordonnées GPS du foyer afin d'identifier par la suite l'ensemble des ouvrages électriques se trouvant à proximité du foyer.

Pour caractériser les expositions, des modèles de régression non paramétriques ont été réalisés. Ils ont permis d'identifier parmi les données recueillies, les facteurs favorisant les expositions moyennes les plus élevées. Chaque série de mesures de champ magnétique a ensuite été décrite par des indicateurs et une classification ascendante hiérarchique a été appliquée sur ces derniers. Trois classes d'exposition ont ainsi été définies. Une régression logistique a été réalisée pour identifier les facteurs favorisant la probabilité d'appartenir à la classe regroupant les individus les plus exposés.

Abstract :

The magnetic fields (MF) at extremely low frequency (ELF) have been suspected, for around 30 years, to be responsible for several pathologies in humans, more precisely, childhood leukemia. The last collective assessment by international expert groups (WHO 2007, SCENHIR 2009) concluded that the last major questioning concerning ELF MF is the statistic correlation observed in several meta-analysis between the increase of childhood leukemia risk and a MF exposure higher than 0.4μT in means over 24h (Ahlbom et al., 2000), without any causal relation.

To study the exposure of the French population to 50 Hz MF, two representative samples of this population (1000 adults and 1000 children) were created using random selection method. Each volunteer wore an EMDEX II measuring and recording MF during 24-hour period and progressively completed a timetable and a questionnaire containing specific information about himself and his home. When returning the measurement device, the pollster noted GPS coordinates at the front door of volunteer home, in order to identify afterwards electric networks near the home.

To characterize mean exposures, nonparametric regression models were applied. They have allowed identifying among the data collected, the factors favoring a higher mean exposure. Each series of MF was then described by indicators and hierarchical clustering was applied. Three classes of exposure were defined. A logistic regression was used to identify factors favoring probability of belonging to the most exposed classes.

Chapitre 1

Introduction générale

1.1 Introduction

En 1979, une étude épidémiologique réalisée par les américains Wertheimer et Leeper a observé, pour la première fois, l'existence d'une association entre des cas de leucémie infantile et certaines caractéristiques du réseau électrique autour des logements des enfants atteints [1]. Selon cette analyse, le fait de résider à proximité d'une ligne de transport électrique pourrait être associé à une augmentation du risque de leucémie infantile. À partir de cette étude la communauté scientifique s'est posée des questions sur les effets des champs magnétiques d'extrêmement basse fréquence (Extremely Low frequency, ELF) sur l'homme. On nomme ELF, les fréquences comprises entre 0 et 300 Hz. Parmi les sources générant des champs magnétiques ELF, on peut dénombrer les lignes de transport d'électricité (leur fréquence est de 50 Hz en Europe et 60 Hz dans les pays américains) et plus généralement tout système utilisant l'électricité distribuée par le réseau électrique 50/ 60 Hz.

Aujourd'hui, les connaissances scientifiques concernant les effets des champs magnétiques ELF sur la santé sont substantielles et fondées sur un grand nombre d'études épidémiologiques, d'études sur l'animal et sur les cellules en culture. De nombreux effets potentiels sur la santé allant des anomalies de la reproduction aux maladies cardio-vasculaires et neurodégénératives ont été examinés mais la question qui reste posée à ce jour concerne la leucémie infantile [2]. En 2001, un groupe de travail du Centre International de Recherche sur le Cancer (CIRC) de l'Organisation Mondiale de la Santé (OMS), composé d'experts scientifiques, a examiné les études se reportant à la cancérogénicité des champs magnétiques ELF et les ont classés dans la catégorie IIB [3]. Ils sont ainsi qualifiés de *peut-être cancérigènes* pour l'homme, pour la leucémie chez l'enfant et pour les expositions supérieures à 0,4 μT en moyenne sur 24 heures [3] (le CIRC n'a pas précisé le type de moyenne). La catégorie IIB est une rubrique utilisée pour qualifier un agent pour lequel on dispose de données limitées concernant sa cancérogénicité chez l'homme et chez les animaux de laboratoire ou pour lequel on dispose d'indications insuffisantes de cancérogénicité chez l'homme mais de peu de preuves de cancérogénicité chez l'animal de laboratoire. La classification des champs magnétiques (CM) ELF par le CIRC est fondée sur les études épidémiologiques réalisées sur la leucémie chez l'enfant, sans qu'un lien de causalité n'ait été démontré par des études expérimentales. Des meta-analyses récentes ont montré une association statistique entre le risque de leucémie de l'enfant et une exposition au CM supérieure à 0,4 μT en moyenne sur 24 heures (à noter que l'OMS ne spécifie pas le type de moyenne à prendre en compte) [2, 14].

Actuellement l'exposition de la population française aux champs magnétiques ELF n'est connue que de manière très approximative. Une étude réalisée

dans le département de la Côte-d'Or sur 240 logements situés à proximité des lignes à haute et très haute tension (HT et THT) a permis d'évaluer les expositions à l'intérieur de ces logements [4]. Mais, d'une part, il s'agit d'un faible échantillon compte tenu de la diversité du parc de logement en France, et d'autre part, il s'agit d'une exposition du logement et non des personnes. De plus, les lignes de transport d'électricité ne sont pas la seule source d'exposition possible même si l'habitat se trouve à proximité de ces ouvrages. En effet tout sujet est exposé à de nombreuses sources de CM du fait qu'il ne reste pas chez lui 24 heures sur 24. Les transports peuvent, en particulier, représenter des sources d'exposition significatives, mais on peut aussi se demander si d'autres lieux de vie ne pourraient pas représenter des sources, comme un lieu de travail, un terrain de sport, un centre commercial, une école, etc. L'exposition peut aussi dépendre des activités, qu'elles soient professionnelles ou non (travail sur ordinateur, repassage, etc) ou de la proximité par rapport à d'autres ouvrages électriques (postes électriques, lignes souterraines, lignes de chemin de fer).

Dans le but de connaître les niveaux d'exposition moyens les plus élevés et d'identifier les sources, les lieux et les activités liés à ces niveaux, le Conseil Supérieur d'Hygiène Publique de France (CSHPF) a recommandé à la Direction Générale de la Santé (DGS) de réaliser une étude d'estimation et de caractérisation des expositions de la population aux champs magnétiques ELF. La DGS a ensuite mandaté l'École Supérieure d'Électricité (SUPELEC) pour faire cette étude. Connue sous le nom d'étude EXPERS (pour EXposition de la PERSonne), elle est réalisée en collaboration avec EDF R&D, le Service des Études Médicales (SEM) de EDF, le gestionnaire du Réseau de Transport d'Électricité de France (RTE) et l'institut de sondage MV2 Conseil.

Le but de cette étude est de quantifier les expositions moyennes de la population (enfants et adultes) (et pas uniquement les champs magnétiques dans les foyers situés à proximité des lignes à haute ou très haute tension) et d'identifier les sources pouvant les influencer. Pour cela, il a été décidé que 1 000 enfants et 1 000 adultes, choisis aléatoirement et uniformement vis à vis de la répartition des ménages dans les régions de France, participent à une enquête nationale d'enregistrement du champ magnétique. En même temps que la mesure de champ magnétique, il leur est demandé de remplir un emploi du temps et un questionnaire. L'étude repose essentiellement sur deux problématiques dont l'une, est la collecte de l'ensemble des informations nécessaires (constitution de l'échantillon, réalisation des mesures, information sur tout ce qui peut influencer l'exposition, etc.). La seconde est liée à l'analyse statistique des informations recueillies (pertinence des mesures de champ magnétique enregistré et des emplois du temps pour chaque personne sondée, recherches des facteurs d'exposition, etc).

1.2 Organisation du mémoire

La thèse est composée de 7 chapitres avec l'introduction.

1. Le chapitre 1 est l'introduction de la thèse.
   Nous soulignons principalement les objectifs de la thèse.
2. Dans le chapitre 2, nous présentons un état de l'art sur les études épidémiologiques et les études d'exposition. Les différentes problématiques rencontrées pendant la thèse sont explicitées.
3. Le protocole de recueil des données est exposé dans le chapitre 3. Une analyse descriptive des données recueillies avec ce protocole est réalisée.
4. Dans le chapitre 4, nous réalisons une analyse descriptive des expositions moyennes. Des tests de comparaison de ces moyennes sont aussi réalisés selon différents critères.
5. Le chapitre 5 est consacré à la caractérisation des expositions moyennes (moyennes arithmétiques et géométriques). Nous réalisons des modèles décrivant ces moyennes à partir des informations recueillies.
6. Nous cherchons, dans ce chapitre, des classes d'exposition après avoir décrit chaque série de champ magnétique par des indicateurs. Nous caractérisons aussi la probabilité d'appartenir aux classes des plus exposés
7. Nous concluons ce travail dans le chapitre 7.

Nous mettons en annexe quelques documents relatifs à l'étude comme la lettre de la Direction Générale de la Santé, un exemplaire d'emploi du temps et un questionnaire.

Chapitre 2

État de l'art et difficultés liées à l'étude

2.1 Introduction

Le nombre de sources de champs électromagnétiques a considérablement augmenté au cours de ces dernières décennies. Chaque individu est exposé en permanence à des champs électriques, magnétiques ou électromagnétiques de toutes natures, qu'ils soient d'origine naturelle ou qu'ils soient directement créés par l'homme pour satisfaire ses besoins en termes de communication, de santé, de transport, de confort, de production, etc. Le problème de l'estimation et de la caractérisation de l'exposition de la population d'un grand pays industrialisé comme la France devient vite complexe, compte tenu de la diversité et de la multiplicité des sources.

Dans ce chapitre, nous présentons un état de l'art des principales études épidémiologiques ou d'exposition. Nous soulignons ensuite les différentes problématiques qui sont liées à cette étude, auxquelles nous souhaitons répondre. Une des ces problématiques est la construction de l'échantillon représentant la population. Une autre est la réalisation des mesures des champs magnétiques (CM).

2.2 Le champ magnétique

En physique, le champ magnétique est une grandeur caractérisée par la donnée d'une intensité et d'une direction, définie en tout point de l'espace, et déterminée par la position et l'orientation d'aimants, d'électroaimants et le déplacement de charges électriques. La présence de ce champ se traduit par l'existence d'une force agissant sur les charges électriques en mouvement.
La force magnétique a été observée à l'origine en constatant que la terre a une force capable d'orienter des aiguilles aimantées. Plus tard (vers 1820), le physicien danois Oersted a mis en évidence le fait qu'un fil électrique ou une bobine de fil électrique parcouru par un courant génère un champ magnétique, orientant ainsi l'aiguille d'une boussole perpendiculaire au fil.

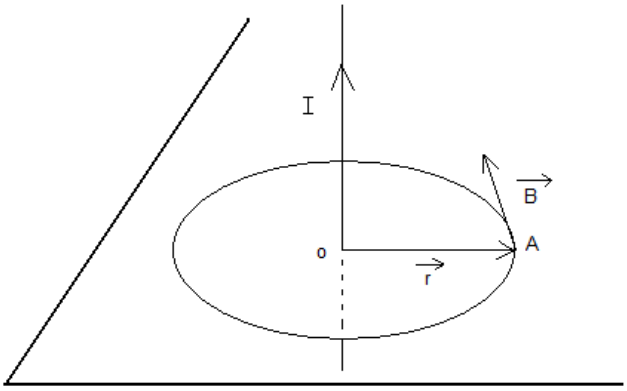

FIG. 2.1 – CM généré par un fil de longueur infinie et parcouru par un courant d'intensité I.

Des expériences ont mis en évidence qu'un long fil conducteur rectiligne transportant un courant stationnaire d'intensité I produit un champ magnétique $\overrightarrow{B}$ qui est tangent à n'importe quel contour C situé dans un plan perpendiculaire à la direction du fil (figure 2.1). La loi d'Ampère lie le courant électrique à son effet magnétique par la relation (2.1).

$$\oint_C \overrightarrow{B}.\overrightarrow{dl} = \mu_0 I. \tag{2.1}$$

où I est le courant final qui passe à travers une surface bornée par le contour C, $\overrightarrow{dl}$ un élément infinitésimal de déplacement le long du contour C et μ_0 la constante universelle ou perméabilité magnétique du vide. Elle a pour valeur $\mu_0 = 4 \times 10^{-7}$ Henry par mètre (H/m).
Dans ces mêmes conditions, on calcule la densité magnétique créée par un fil en un point A, situé dans un plan normal au fil à la distance r de celui-ci par la relation (2.2). Cette relation montre que l'intensité du CM généré par un fil rectiligne parcouru par un courant décroît en $1/r$.

$$B = \frac{I}{2\pi r}\mu_0. \tag{2.2}$$

avec I en Ampère (A), r en mètre (m) et B en Tesla (T).

Dans toute la suite, la densité magnétique sera appelée champ magnétique. Grâce à des méthodes de calcul, plus ou moins sophistiquées, on est en mesure de quantifier le champ magnétique à proximité des ouvrages sous tension parcourus par un courant électrique. Cette intensité dépend de la

disposition géométrique des conducteurs parcourus par le courant et est proportionnelle à l'intensité du courant.

Il existe deux types de champs magnétiques :

1. **Les champs magnétiques naturels**
   Il existe un champ magnétique naturel statique à la surface de la terre. La terre se comporte comme un aimant. Ce phénomène est dû aux mouvements du noyau métallique liquide des couches profondes de la terre. L'intensité de ce champ statique est comprise entre 35 et 50 μT selon la latitude. Il peut toutefois être perturbé localement par des composés ferreux. L'homme, comme tous les êtres vivants est sous influence du champ magnétique terrestre. Les éclairs et la lumière produisent aussi des champs électromagnétiques de l'ordre de 10^6 Hertz (Hz) pour la foudre et autour de 10^{14} Hz pour la lumière.
2. **Les champs magnétiques artificiels**
   Il existe différentes catégories de CM artificiels. Ces différences sont relatives à la gamme de fréquence des champs (figure 2.2). Dans cet ensemble, se trouvent les champs magnétiques générés par la production, le transport et l'utilisation de l'électricité. Leur fréquence est en Europe de 50 et 60 Hz en Amérique du Nord. Elle est beaucoup plus basse que celle des ondes radio (autour de 100 MHz) et que celle de la lumière visible (autour de 10^{14} Hz). Au-delà, commence la gamme des champs électromagnétiques dits ionisants (on parle de rayonnement non ionisant pour les champs de fréquence inférieure à la lumière visible). Leur énergie est suffisante pour rompre les molécules et ioniser les atomes (les ultras violets, les rayons X, etc.) alors que celle des champs 50/ 60 Hz est très faible [2].

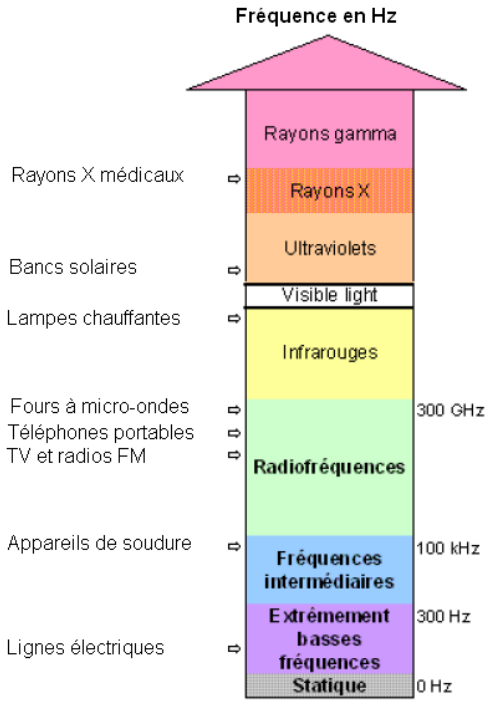

FIG. 2.2 – Illustration des différents types de champs magnétiques en fonction de la fréquence et des sources potentielles [5].

Depuis les années 70, la question d'une possible nocivité du champ magnétique 50/ 60 Hz a été posée dans la communauté scientifique. Des projets de recherches ont été mis en place et des études ont été réalisées.

2.3 État de l'art

Les articles de la littérature abordent deux aspects :

1. les aspects épidémiologiques,
2. les aspects liés à la mesure de l'exposition.

2.3.1 Epidémiologie

Pour mesurer la force de l'association entre une exposition et la survenue d'un événement dans les études de cas-témoin, on utilise l'odds ratio (OR) ou « rapport des cotes ». C'est une approximation du risque relatif. Si on note $\tilde{p}_0$ la probabilité d'avoir l'événement considéré dans le groupe des cas et $\tilde{p}_1$ celle d'avoir cet événement dans le groupe des témoins, l'OR est définie par :

$$OR = \frac{\tilde{p}_0(1-\tilde{p}_1)}{\tilde{p}_1(1-\tilde{p}_0)}$$

S'il est proche de 1, l'événement est indépendant du groupe et s'il est supérieur à 1, l'événement est plus fréquent dans le groupe des cas que dans le groupe des témoins.

Dans ce domaine, la première étude est réalisée en 1979 par Wertheimer et Leeper [1]. Cette étude réalisée à Denver (Colorado, États-Unis) sur 355 cas de cancer de l'enfant et 344 témoins a montré une association statistique entre « le code de câblage » du foyer de résidence et le risque de leucémie. Pour évaluer le niveau d'exposition aux CM, Wertheimer et Leeper ont supposé que la source principale d'exposition était les lignes électriques avoisinantes et qu'une inspection visuelle de ces lignes leur permettrait d'évaluer la charge électrique transportée. À partir de ce paramètre et de la distance séparant la ligne du domicile, ils ont classé les logements en quatre groupes :

- VHCC : Very High Current Configuration
- OHCC : Ordinary High Current Configuration
- OLCC : Ordinary Low Current Configuration
- VLCC : Very Low Current Configuration

en fonction de l'intensité estimée des CM et de la distance par rapport à la résidence. Ils ont qualifié leur système d'inspection visuelle de « code ». Ce code est maintenant connu sous le nom de « code Wertheimer Leeper » ou simplement de « code de câblage ». La comparaison des logements, où avaient vécu les enfants décédés de cancer, avec celle des logements des témoins a montré que le risque est 2,98 (IC=[1,72 ; 5,15]) fois plus important en cas de HCC par rapport au LCC.
Le codage de câblage permet de classer un logement en fonction de sa distance à un type particulier de source de CM. Cette approche ne considère que l'exposition aux champs magnétiques générés par le système de distribution « visible » de l'électricité. D'autres paramètres comme les lignes souterraines, le câblage de l'habitation, ou l'ensemble des appareils domestiques ne sont pas pris en compte [1]. D'autres études ont suivi, la majorité s'est limitée aux foyers à proximité des lignes de transport d'électricité. Les estimations des CM sont effectuées soit en utilisant les caractéristiques des lignes électriques soit en utilisant l'historique de la charge électrique de la ligne [6, 7, 8, 9].

Les dernières études ont cherché à quantifier l'exposition par des mesures réelles effectuées à point fixe (dans les lieux d'habitation ou les lieux de travail) [11]. Ces mesures fournissent l'exposition en un lieu donné, à partir de laquelle on estime l'exposition des personnes concernées en se fondant sur des hypothèses de durée de séjour en tel ou tel lieu. Plus récemment, des mesures directes de l'exposition des personnes au moyen d'enregistreurs portables ont été utilisées [19]. Ces mesures peuvent porter sur des durées

variables (de quelques heures à une semaine) [19].

En 1997, une autre étude est réalisée dans 9 États américains par Linet [12]. Elle évalue l'association entre la leucémie chez l'enfant et l'exposition au CM en milieu résidentiel. Seule la leucémie lymphoblastique aiguë (LLA) est étudiée. Cette étude a une bonne taille de l'échantillon (638 cas et 620 témoins), un bon taux de participation pour la mesure du CM dans les résidences des cas. Les résultats montrent de faibles associations statistiquement non significatives sauf lorsque l'on considere l'exposition se situant entre 0,40 et 0,49 μT dans la residence. L'odds ratio est alors de 3,28 (IC = [1,15 ; 9,39]). Il faut néanmoins préciser que la taille d'échantillon pour les expositions au CM supérieures à 0,2 μT est faible. Des résultats non significatifs sont observés lors de l'utilisation de la codification du reseau électrique selon les methodes de Wertheimer-Leeper. Les auteurs concluent que leur étude apporte peu de preuves de l'existence d'un lien entre les niveaux élevés de CM mesurés (moyennes pondérées sur le temps) ou estimés selon la codification du reseau électrique (VHCC) et le risque de leucémie lymphoblastique aiguë.

Les études les plus récentes ont porté sur un grand nombre de cas et ont évalué l'exposition des enfants à partir de mesures de champs magnétiques ambiants dans les résidences ou dans les établissements scolaires [13]. Les études les plus importantes offrent peu ou pas d'indication de risque accru de cancer chez les enfants [13]. Par exemple, une étude réalisée au Royaume Uni sur 3 838 cas et 7 629 témoins a montré que le risque de développement de leucémie pour les enfants ayant une exposition moyenne supérieure à 0,2 μT était de 0,90 (IC=[0,49 ; 1,63]) par rapport aux enfants exposés à un CM moyen inférieur à 0,1 μT [13]. Pour les expositions supérieures ou égales à 0,3 μT, le risque relatif était de 0,93 (IC=[0,30 ; 2,91]). Pour les expositions supérieures ou égales à 0,4 μT, le risque relatif n'est pas donné car l'intervalle de confiance est trop large et les auteurs disent que le résultat n'est pas fiable. Cette étude est connue sous le nom de « étude UKCCS » pour « United Kingdom Childhood Cancer Study ».

Une méta-analyse réalisée par Ahlbom [14], regroupant des données de neuf études entreprises en Europe, en Amérique du Nord et en Nouvelle-Zélande, et portant sur 3 203 cas de leucémie et 10 338 témoins, a conclu à l'absence d'une association entre la leucémie et un CM résidentiel de 0,1 μT à 0,4 μT en moyenne sur 24 heures, par rapport au groupe de référence dont l'exposition était inférieure à 0,1 μT. Cette même analyse a montré qu'un niveau de CM résidentiel supérieur à 0,4 μT en moyenne géométrique sur 24 heures était associé à un risque relatif de 2,00 (IC=[1,27 ; 3,13]), bien que des biais de sélection puissent expliquer une partie de cette augmentation. En fait d'une part, les individus sont sélectionnés dans différents pays avec

des protocoles non nécessairement identiques, d'autre part, les niveaux d'exposition ne sont pas des expositions personnelles mais des enregistrements réalisés aux domiciles des sujets. Ceci peut induire un biais de mesure. C'est cette étude qui a amené le Centre International de Recherche sur le Cancer (CIRC) à classer ces champs dans la catégorie *peut-être cancérigène* pour l'homme.

En 2005, une autre étude épidémiologique réalisée par Draper est publiée [15]. Elle avait pour objectif de déterminer s'il y a une association entre la distance de résidence à la naissance par rapport aux lignes hautes tensions de transport de l'électricité (132 et 400 kV) et l'incidence de la leucémie et d'autres cancers infantiles en Angleterre et au Pays de Galles (toutes les lignes de 132 kV ne sont pas prises en compte). Cette étude de cas a inclu 29 081 enfants atteints de cancer dont 9 700 sont atteints de leucémie. Tous les enfants ont moins de 14 ans et sont nés en Angleterre ou au Pays de Galles entre 1962 et 1995. Chaque sujet était individuellement apparié avec un temoin de même sexe, de date de naissance et de zone d'enregistrement proches.
Les résultats ont montré que, par rapport aux enfants qui vivent à plus de 600 m d'une ligne à la naissance, les enfants vivant à moins de 200 m ont un risque relatif de leucémie de 1,69 (IC=[1,13 ; 2,53]). Pour ceux nés à des distances comprises entre 200 et 600 m, le risque relatif est de 1,23 (IC=[1,02 ; 1,49]). Ils ont montré l'existence d'une variation significative du risque avec l'inverse de la distance à la ligne. Mais cette relation n'est plus observée avec l'inverse du carré de la distance, pourtant plus proche de la réalité physique en terme de CM. D'ailleurs les auteurs restent prudents dans leur conclusion : « il est surprenant de retrouver des effets aussi loin des lignes ». À 200 m, le CM du à la ligne est inférieur au CM moyen du aux autres sources au domicile. Les auteurs soulignent l'incertitude pour savoir si cette relation statistique représente une relation causale. Il n'y a eu aucune mesure du CM. Aucun excès de risque en lien avec la proximité des lignes n'a été trouvé pour les autres cancers infantiles.

L'étude la plus récente est celle réalisée au Japon par Kabuto et ses collaborateurs [16]. Cette étude cas-témoin est réalisée sur la population générale. Les auteurs ont analysé 312 enfants de 0 à 15 ans dont on a diagnostiqué une leucémie lymphoblastique aiguë (LLA) ou une leucémie myéloïde aiguë (LMA) entre 1999 et 2001 et 603 enfants témoins présentant des caractéristiques similaires en termes de genre, d'âge et le lieu de résidence. Les niveaux moyens hebdomadaires de CM ont été mésurés dans la chambre des enfants. Les mesures de CM dans les groupes cas et témoins ont été réalisées à la même période afin de tenir compte des variations saisonnières. L'association a été évaluée sur base de modèles de régression logistique conditionnels. Les odds ratio pour les enfants ayant des niveaux de CM égaux ou supérieurs

à 0,4 μT comparés à ceux de la catégorie de référence (CM plus petit que 0,1 μT) étaient de 2,6 (IC=[0,76 ; 8,60]) pour LMA et LLA et 4,7 (IC=[1,15 ; 19,0]) pour LLA seul. Les résultats n'ont pas été modifiés par la prise en compte de facteurs de confusion. La plupart des cas de leucémie dans la catégorie d'exposition la plus élevée était exposée à des niveaux bien plus importants que 0,4 μT : cette relation porte sur 6 cas et 3 témoins.

2.3.2 Mesure d'exposition

Les mesures d'exposition sont le plus souvent incluses dans des études épidémiologiques. L'exposition de la population aux CM résulte de multiples sources, que ce soit en milieu résidentiel, au travail ou lors de la fréquentation de lieux publics. L'intensité des CM varie en fonction du type de source émettrice et selon la distance entre l'individu et la source. Nous distinguons, dans ce paragraphe, les méthodes de mesure de l'exposition et les résultats des études d'exposition.

2.3.2.1 Les méthodes de mesure de l'exposition

Plusieurs études d'exposition sont réalisées avec des méthodes totalement différentes.

Aux États Unis, Linet a mesuré avec un EMDEX-C les champs magnétiques pendant 24 heures dans les chambres d'enfants. L'appareil de mesure est posé sous ou juste à côté du lit de l'enfant. D'autres mesures de 30 secondes sont aussi réalisées dans la chambre des parents, dans la chambre où dormait la mère pendant la grossesse, à la cuisine et à 0,9 m de la porte d'entrée de la maison à l'extérieur [12].

Pour l'étude UKCCS, les mesures sont réalisées aux domiciles des enfants et dans leurs établissements scolaires. Au domicile, il y a eu deux phases de mesures :

- La première phase est réalisée en trois étapes dans l'ordre suivant :
  1. 3 minutes de mesures au centre de la chambre à coucher, au centre du lit et au centre de l'oreiller se trouvant sur le lit de l'enfant.
  2. 90 minutes de mesures au centre de la chambre à coucher des parents.
  3. Reprendre l'étape 1 après avoir réalisé les mesures dans la chambre des parents.

  L'appareil de mesure (EMDEX II) était accroché dans un piquet à une hauteur de 1 m du plancher et d'au moins à 1 m de tout appareil en fonctionnement. Pour cette phase, des informations relatives au temps de sommeil de l'enfant et au temps passé à l'école sont demandées.

- La seconde phase est conditionnée à la première. Dans la première phase, lorsque des expositions liées à tout type de chauffage sont identifiées, les mesures de la seconde sont réalisées en période hivernale, convenue avec la compagnie d'électricité.
    1. 4 spots de 3 minutes de mesures dans la chambre à coucher des parents, à côté du lit, au centre du lit et au centre de l'oreiller de l'enfant.
    2. 48 heures de mesures réalisées à côté du milieu du lit de l'enfant.
    3. Reprendre les quatre spots de la première étape après l'étape 2.

Dans les établissement scolaires, les mesures sont réalisées pendant que les systèmes de chauffage fonctionnaient normalement (entre octobre et mars). Elles sont menées en deux étapes :

1. La première concerne les enfants ayant occupé une seule salle de classe pour la majorité du temps passé à l'école au cours de la période hivernale de référence (généralement dans les écoles primaires). Pour cette catégorie, les mesures sont effectuées en cinq spots de 2 minutes à proximité du centre et des quatre angles de la classe.
2. Lors plusieurs salles de classe ont été utilisées (généralement dans les écoles secondaires), des mesures ponctuelles ont été effectuées jusque dans cinq salles (les plus fréquemment utilisées au cours de la période hivernale). Dans chacune de ces salles, des mesures sont effectuées à proximité du centre de chaque salle. Le temps total de mesure est de 10 minutes et les durées de mesures sont égales pour les différentes salles.

Au Canada, Green [17] a utilisé trois méthodes pour évaluer l'exposition :

- Code de câblage (3 versions).
- Mesures en différents points de la résidence : dans la chambre à 30 cm au dessus du centre du lit, et dans deux autres pièces les plus fréquemment utilisées par l'enfant, généralement le séjour et la cuisine. Ces mesures étaient faites par l'enquêteur (la moitié ont été répliquées par un technicien). Une moyenne temporelle était ensuite faite avec ces 3 points de mesures.
- Appareil de mesure personnelle (Positron) porté par l'enfant, réglé pour faire une mesure toutes les minutes pendant 2 jours. Un sac à dos était fourni pour les jeunes enfants, une pochette en bandoulière pour les plus grands, et on demandait à la mère de garder le Positron près de l'enfant pour ceux qui n'étaient pas en âge de le porter. Sur le questionnaire d'emploi du temps, on demandait de préciser quand l'appareil n'était pas porté mais était près de l'enfant. La nuit, le Positron était posé près du lit à un emplacement choisi par l'enquêteur de

manière à éviter les appareils électriques et à approcher l'exposition sur le lit de l'enfant.

Toujours au Canada, McBride [10] a évalué l'exposition au champ magnétique par :

- code de câblage (2 versions)
- appareil de mesure personnelle (Positron) porté par l'enfant pendant 2 jours dans un petit sac à dos
- mesure pendant 24h dans la chambre de l'enfant avec un Positron. L'emplacement était prédéfini par l'enquêteur, mais n'est pas précisé dans la publication.

En Allemagne, Schüz [18] a également évalué l'exposition en des points fixes du domicile :

- Mesure pendant 24 heures toutes les secondes sous le matelas de l'enfant avec un appareil FW2a (50 Hz et 16 Hz).
- Mesure pendant 24 heures dans la pièce où l'enfant passe le plus de temps en dehors de sa chambre avec un EMDEX II, à un emplacement fixe (loin de tout appareil électrique).

Pour l'étude de Kabuto [16], les niveaux moyens hebdomadaires de CM ont été déterminés à point fixe dans la chambre des enfants. Ces mesures sont basées sur les résultats de Friedman [19] :

- Mesures toutes les 30s pendant 1 semaine dans la chambre de l'enfant avec un EMDEX LITE (40 Hz - 1 kHz) (pas de précision dans la publication sur le point exact de mesure)
- Mesures pendant 5 minutes avec un EMDEX II (40-800 Hz) en plusieurs points à l'intérieur et à l'extérieur de la maison.

En conclusion, les études d'exposition sont majoritairement basées sur une estimation des CM dans des endroits fixes sauf celles de Mc Bride et Green où les individus ont porté un appareil mesurant leur exposition personnelle. Le plus souvent, les auteurs ont fait attention à ce que les mesures ne soient pas liées à tout appareil électrique.

2.3.2.2 Résultats des études d'exposition

En France, une étude a été réalisée dans la région de Côte-d'Or sur 240 foyers situés à proximité des lignes à haute tension et très haute tension. Elle a montré que l'intensité moyenne à l'intérieur de ces foyers était de 0,05 μT [4]. Les facteurs d'exposition identifiés sont la présence de lignes de transport d'électricité (distance par rapport au foyer et le type de la ligne ou la tension) et l'année de construction du bâtiment. Cette étude donne une estimation de l'exposition dans des endroits fixes du foyer (au salon, dans la

chambre à coucher) et non celle des habitants. Même au domicile, le fait de supposer que les résidents sont exposés au CM moyen enregistré dans leur foyer conduirait à un important biais car l'exposition dépend de plusieurs facteurs. Elle varie avec le temps, l'espace, la proximité de l'individu à l'ensemble des appareils électriques sous tension, etc.

Le California EMF Program a mené durant 3 ans un vaste programme d'évaluation des niveaux de CM en milieu scolaire. Les mesures ont été réalisées sur 5403 locaux dont 3193 salles de classe reparties dans 89 écoles. Cette étude a montré que la proportion de classes ayant une moyenne de CM supérieure à 0,3 μT est de $2,1\%$ tandis qu'elle est de $1,2\%$ pour un seuil de 0,4 μT.

Deadman et ses collègues [20] ont mesuré à l'aide d'appareils Positron, dans le cadre d'une étude épidémiologique, l'exposition aux CM d'enfants de 5 provinces du Canada. Dans l'étude, 214 enfants ont des mesures en milieu scolaire. Parmi ces 214 enfants, 92 résident au Québec. La moyenne arithmétique du CM pendant la période scolaire pour l'ensemble des enfants canadiens est de 0,12 μT. Elle est de 0,14 μT pour les enfants québéquois.

Au Québec, Gauvin et ses collaborateurs [21] ont mesuré l'intensité du CM dans 21 classes de 10 écoles pourvues de planchers électriques chauffants. Les mesures ponctuelles prises à 50 cm de hauteur montrent des niveaux élevés pour trois locaux où les systèmes de chauffage sont munis d'un dispositif de transformation de tension. Les moyennes de CM dans ces locaux se situent entre 31 et 39 μT comparativement à des moyennes entre 0,05 et 2,4 μT pour les planchers chauffants sans dispositif de transformation de tension. Des mesures prises au niveau du sol montrent, pour les systèmes avec transformation de tension, des dépassements de la recommandation de l'ICNIRP avec une intensité maximale de près de 500 μT.

Pour l'étude UKCCS [13], moins de $0,4\%$ des enfants ont observé un CM moyen supérieur à 0,4 μT. Pour la méta-analyse d'Ahlbom [14], les champs magnétiques au-delà de 0,4 μT étaient rares : seulement $0,8\%$ des sujets de l'étude étaient exposés à un champ résidentiel moyen, supérieur ou égal à ce niveau.

2.4 Réglementations

2.4.1 Réglementation de l'ICNIRP

À partir des données sur les effets des CM sur la santé, la communauté scientifique a adopté des recommandations. À 50/60 Hz, les effets avérés des champs sont une stimulation possible du système nerveux pour des champs

très élevés en valeur instantanée. L'ICNIRP (International Commission on Non-Ionizing Radiation Protection) et l'IEEE (Institute of Electrical and Electronics Engineers) ont donc fixé des valeurs pour se protéger de ces effets. L'ICNIRP a donné un niveau de référence de 100 μT pour le public et 500 μT pour les travailleurs [22]. Les restrictions de base de l'ICNIRP sont exprimées en termes de densité de courant induit dans le système nerveux central alors que celles de l'IEEE sont exprimées en champ électrique induit. En 1999, l'Union Européenne a adopté cette recommandation pour l'exposition du public et en 2004, une directive sur l'exposition professionnelle, fondée sur les recommandations de l'ICNIRP, a été adoptée [23, 24]. D'autres recommandations sont aussi publiées en 2002 par IEEE. Elles spécifient, pour les champs de 60 Hz, des « Maximum Permissible Exposure » de 904 μT pour le public et 2 710 μT pour les travailleurs [25].

2.4.2 Autres politiques

L'OMS recommande de suivre les recommandations de l'ICNIRP. La plupart des états l'ont fait. La France se réfère à la recommandation européenne de juillet 1999 [23] sans restriction particulière. Mais récemment certains états ont ensuite mis en place des politiques de précaution pour rassurer la population. Kheifets [26, 27] a récemment publié 2 articles dans lesquels sont présentés quelques exemples de réglementations adoptées sur la base de diverses politiques de précaution dont :

- L'Italie a adopté divers critères (ou mesures) à respecter, ceux-ci étant définis principalement en fonction des infrastructures localisées à proximité des installations électriques. Elle recommande comme limite d'exposition pour les champs magnétiques dans le cas des lignes à haute tension des niveaux à ne pas depasser de 100 μT. Par mesure de prudence et afin de se protéger des effets possibles à long terme, elle a adopté une *attention value* de 10 μT (médiane mesurée sur 24 heures dans des conditions normales d'opération) pour les jardins d'enfants, les maisons résidentielles, les locaux scolaires et les aires où les personnes peuvent rester 4 heures ou plus par jour. Également, lors de la conception des nouvelles lignes électriques au voisinage des mêmes emplacements précités, l'Italie a adopté un *quality goal* de 3 μT afin de minimiser progressivement l'exposition aux CM générés par les lignes électriques à 50 Hz (The President of the Council of Ministers, Italy, 2003).

- La Suisse a également adopté des limites spécifiques de CM (*installation limit values*) qui doivent être respectées à certains endroits comme les appartements, les écoles, les hôpitaux, les endroits permanents de travail et les jardins d'enfants. Pour ces installations, la valeur limite

retenue a été fixée à 1 μT (Swiss Federal Council, 2000).

– En Israël, en 2001, le ministère de l'Environnement a retenu une valeur limite de 1 μT à ne pas excéder dans les aires publiques (mesure basée sur une exposition moyenne de 24 heures (Israel Ministry of the Environment, 2005).

– D'autres pays ont adopté des mesures visant à limiter la construction de nouveaux sites à proximité des sources d'exposition. L'Irlande n'accorde pas de permis de construction aux compagnies d'électricité à proximité des écoles et garderies (à moins de 22 mètres). Aux Pays-Bas, les nouvelles écoles doivent se trouver à une certaine distance afin que l'exposition des enfants n'atteigne pas 0,4 μT en moyenne.

Toutes ces restrictions sont basées sur des mesures de précaution du public dans le cadre de la gestion politique de risque. Elles ne sont en aucun cas des seuils de nocivité. L'ensemble des études réalisées sur les effets des CM ELF sur l'homme n'ont pas montré d'effets nocifs sur la santé. Celles qui ont montré un risque accru de développement de leucémie infantile ne sont pas reproductibles (on peut donc imaginer qu'il n'y ait pas une relation de causalité). Le risque pourrait être lié à un agent méconnu.

2.5 Principales difficultés

En général, pour étudier un caractère spécifique dans une population, on se sert d'un échantillon représentatif de cette population. En effet, il est pratiquement impossible de réaliser l'étude sur toute la population lorsqu'elle est de grande taille. C'est ce qui se passe dans les enquêtes d'opinion ou de consommation. La manière de choisir cet échantillon est propre à chaque situation.

2.5.1 Choix des individus

Les études épidémiologiques relatives aux CM ELF sont principalement fondées sur une estimation des expositions à partir des CM enregistrés dans les domiciles des sujets. Cette démarche ne permet pas de mesurer l'exposition exacte des personnes concernées. Elle est le plus souvent basée sur des individus spécifiques. Notre étude est très différente de ces dernières car on ne cherche pas à connaître les effets des CM ELF sur la santé mais à savoir comment la population est exposée à ces champs, en termes de champ moyen. Les difficultés liées à cette étude sont de différentes natures. La plus importante est le choix des individus pour que l'échantillon représente la population avec toute sa diversité. Une autre difficulté est le recueil des

informations nécessaires pour identifier les facteurs qui favorisent une exposition moyenne journalière. Pour répondre à ces questions, il a été décidé de confier la sélection des individus et la collection de toutes les informations nécessaires à un institut de sondage. Pour cela, un appel d'offre a été lancé et le groupe MV2 Conseil a été choisi pour construire cette base de données. Cette dernière doit être composée des champs magnétiques enregistrés pendant 24 heures (au moins) par un échantillon de 1 000 enfants (de moins de 15 ans) et 1 000 personnes dites adultes (de 15 ans et plus) ainsi que l'ensemble des informations permettant d'identifier les sources des champs. L'échantillon doit être réparti sur toute la France avec un critère de proportion sur la répartition des ménages par région. Enfin, on doit se demander quelles informations recueillir pour pouvoir caractériser les niveaux d'exposition.

2.5.2 Pertinence des données

Une fois les mesures réalisées, se pose la question de la pertinence des champs enregistrés et de l'ensemble des informations. On doit s'assurer en outre de la validité des données avant de les analyser. Cette vérification est longue à réaliser car elle se fait à la main. Cette vérification est basée sur des problèmes techniques tels que :

- Est-ce que l'appareil de mesure a été porté pendant 24 heures ?
- Est-ce que l'emploi du temps et le questionnaire sont bien remplis ?
- Est-ce que la série de CM reflète l'emploi du temps ?

La figure 2.3 est un exemple d'une série (la composante large bande voir paragraphe 3.3.1) découpée à l'aide de l'emploi du temps de la personne concernée. Ce volontaire a observé des pics d'exposition dans les transports ferroviaires (1,6 μT), dans les transports non électriques (1,0 μT) et dans le centre commercial où il faisait ses courses (1,2 μT).

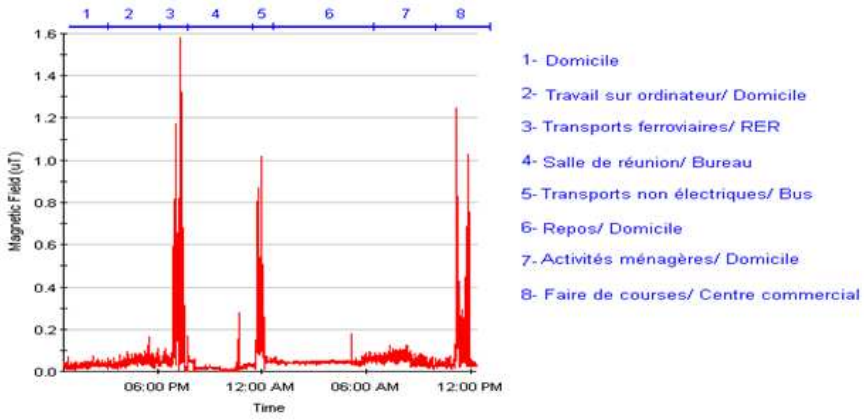

FIG. 2.3 – Parallélisme entre la composante broad band d'une série de CM et l'emploi du temps de la personne concernée.

La vérification de la cohérence entre la série de CM et l'emploi du temps est peut être la plus compliquée. Nous avons ainsi visualisé chaque courbe de CM et vérifié si elle est plausible avec l'emploi du temps (exemple : un CM constant sur 24 heures quand la personne dit avoir eu des activités dans différents lieux n'est pas plausible).

2.5.3 Méthodes statistiques

Une fois les problématiques d'ordre technique résolues, se pose la question du choix des grandeurs qui résument l'exposition d'une personne et des méthodes à utiliser pour bien caractériser l'exposition à partir des informations recueillies. Les méthodes statistiques classiques sont-elles applicables ? A-t-on suffisamment de données pour pouvoir les appliquer ? Les informations ainsi recueillies permettent-elles de bien caractériser les niveaux d'exposition ? Il est clair que l'exposition ne dépend pas seulement du temps et de l'espace mais aussi de l'activité et des ouvrages électriques proches de l'individu sondé. Pour espérer mieux caractériser l'exposition, il faut avoir connaissance de toutes ces informations dans le temps et dans l'espace et procéder éventuellement à une caractérisation par lieu d'activité. En l'absence de toutes les informations nécessaires, les modèles peuvent ne pas produire des modèles réellement prédictifs.

2.6 Conclusion

Dans ce chapitre, nous avons présenté les protocoles et les résultats des études les plus importantes du domaine. La majorité des études épidémiologiques sont basées sur une estimation des CM au domicile. Les résultats des études épidémiologiques sont loin d'être concluants : il n'y a pas de preuves

scientifiques établies pour l'association observée dans certaines études entre l'augmentation du risque de leucémie et les CM 50 Hz. Ce manque d'explication scientifique a conduit certains pays à adopter des politiques de précaution en limitant les niveaux d'exposition en dessous des recommandations internationales. La France applique les recommandations de l'Union Européenne mais veut aujourd'hui savoir comment ses citoyens sont exposés aux CM 50 Hz. Les études d'exposition personnelle les plus récentes ont été réalisées en 1999 au Canada, l'une par Green et l'autre par McBride. Les auteurs ont réalisé des mesures personnelles en faisant porter les appareils de mesure. Cette façon de faire est la plus pertinente lorsqu'on veut connaître l'exposition réelle d'une personne. Elle reste néanmoins la plus difficile à mettre en œuvre lorsqu'on s'intéresse à l'exposition à l'échelle d'un pays. Les problématiques sont de plusieurs niveaux en commençant par la manière de choisir les individus, les appareils à utiliser et les informations à demander. Après peuvent se poser les questions d'ordre technique comme la pertinence des mesures et des informations ou encore les méthodes à utiliser pour expliquer les niveaux d'exposition.

Chapitre 3

Recueil des données

3.1 Introduction

Pour estimer l'exposition de la population aux CM ELF, la première étape consiste à construire un échantillon de cette population, choisi selon des règles rigoureuses, garantissant sa représentativité par rapport à la population nationale. Chaque personne faisant partie de l'échantillon peut ensuite réaliser les mesures. Sous ces conditions, on peut estimer les expositions moyennes de la population une fois la pertinence des séries des CM vérifiée. Pour la caractérisation, il faut renseigner de nombreuses informations relatives à la période de mesure et s'assurer qu'elles sont bien fournies systématiquement.
Dans ce chapitre, nous montrons comment les individus sont sélectionnés et la manière dont les mesures sont réalisées. Une fois présenté l'ensemble des informations demandées, nous analyserons la nature de la base de données ainsi obtenue.

3.2 Sélection des individus

Pour réaliser cette étude, nous devons mettre en place un protocole de sélection des individus pouvant constituer un échantillon représentatif de la population. Il existe deux sortes de méthodes pour créer un échantillon dans une population donnée : la méthode des quotas et celle dite du tirage aléatoire.

3.2.1 Méthode des quotas

Elle est la plus utilisée par les instituts de sondage. Il s'agit de reconstituer une population en miniature, c'est-à-dire de construire un échantillon dans lequel les individus sont répartis selon les mêmes proportions que dans le groupe à étudier. Or il existe bien sûr un grand nombre de caractéristiques possibles (sexe, âge, catégorie socio-professionnelle, lieu de résidence, salarié du public/ privé, etc.). Le nombre de facteurs pris en compte dépend de la question posée et donc de la précision escomptée. Ce nombre est toutefois limité car chacune des catégories doit comprendre un nombre suffisant d'individus. La recherche sociologique tente de cerner avec soin ces facteurs, qui ne font toutefois pas l'unanimité, tandis que dans les études de marketing , les facteurs sont en général moins nombreux et la méthode est facile à mettre en œuvre. En principe, la taille de l'échantillon est indépendante de la taille de la population que l'on souhaite étudier (habitants d'une ville, d'un pays) ; en revanche elle dépend fortement de la marge d'erreur statistique que l'on accepte dans le sondage et de la nature de ce que l'on cherche à identifier. Ainsi, définir des quotas revient à définir une stratification multiple sur la population.
Dans ce type de méthode, l'enquêteur est libre d'interroger qui il veut,

pourvu qu'il respecte les quotas qui lui sont fixés. Dans le cadre de notre étude, mis à part un quota par région, on ne définit aucun quota a priori. Cette méthode n'est donc pas appropriée à notre étude.

3.2.2 Méthode de tirage aléatoire

Le sondage aléatoire consiste à tirer dans une population de taille N un échantillon de taille fixe n, sans remise, à partir des seuls identifiants des individus de façon à ce que chaque individu de la population ait la même probabilité d'inclusion, et cela sans aucune manipulation préalable dans la population ni intervention d'aucune information auxiliaire.
Cette méthode nécessite de disposer d'une base de sondage contenant l'ensemble de toute la population. Il s'agit par exemple, pour un sondage lié aux ménages en France, d'une base des données composée des numéros de téléphone fixe auxquels il faut rajouter les numéros en liste rouge et des portables « *only* » (les ménages qui n'ont pas de téléphone fixe mais uniquement un téléphone portable). La sélection des ménages peut ainsi se baser sur le tirage d'un numéro de téléphone dans cette base de manière aléatoire.

3.2.3 Méthodologie de sélection des individus

Pour réaliser cette étude, il était primordial de trouver un institut de sondage compétent, capable de collecter l'ensemble des informations nécessaires. Le groupe MV2 Conseil (http ://www.mv2group.com) a été choisi pour réaliser cette enquête.

La méthode requise pour sélectionner les individus est la méthode de tirage aléatoire, reconnue de manière internationale pour ce genre d'études. Pour la mettre en place, MV2 Conseil a dû construire une base de sondage contenant l'ensemble des numéros de téléphone (non professionnels) dans laquelle un tirage aléatoire est appliqué.

3.2.3.1 Base de données initiale

La méthode de tirage aléatoire permet à n'importe quel individu résidant sur le territoire national d'être inclus dans l'échantillon soumis aux tests de l'étude. Il s'agissait donc de constituer, dans un premier temps, une base de candidats éligibles ayant tous la même chance d'être sélectionnés dans le protocole de l'étude. Cette phase s'est déroulée en plusieurs étapes :

1. Tirage au sort d'un échantillon de numéros de téléphone représentant la population nationale selon les statistiques de l'INSEE (Institut Nationale de la Statistique et des Etudes Economiques) dans l'annuaire de France Telecom.

2. Création d'un fichier de numéros de téléphone en liste rouge en tirant aléatoirement une base de numéros de téléphone fixe dans l'annuaire de France Telecom puis en rajoutant 1 sur chaque numéro (exemple 01-42-58-96-37 devient 01-42-58-96-38) et en éliminant ceux qui sont présents dans l'annuaire après une recherche inversée (on suppose que les numéros en liste rouge ne sont pas successifs). Les numéros restants sont ensuite chargés dans un automate d'appel qui se charge de vérifier l'existence de la ligne. Les faux numéros sont supprimés. Il reste au final une liste de numéros qui ne figurent pas dans l'annuaire (numéros en liste rouge ou professionnels). Les numéros professionnels sont supprimés lors du questionnaire de recrutement.
3. Génération de numéros de mobile, selon les indicatifs qui existent en France, et vérification par l'automate d'appel de l'existence de la ligne. Les numéros des personnes ayant également un numéro de téléphone fixe sont éliminés. Les numéros restants sont appelés « *mobiles only* ».

Région	Ménages	Pourcentage	Enfants	Adultes
Alsace	678 586	2,85	29	29
Aquitaine	1 212 480	5,09	51	51
Auvergne	556 229	2,34	23	23
Basse Normandie	571 914	2,40	24	24
Bourgogne	670 964	2,82	28	28
Bretagne	1 209 901	5,08	51	51
Centre	999 705	4,20	42	42
Champagne Ardenne	539 949	2,27	23	23
Corse	106 236	0,45	4	4
Franche Comté	452 198	1,90	19	19
Haute Normandie	698 463	2,93	29	29
Île-de-France	4 509 623	18,94	189	189
Languedoc Roussillon	968 616	4,07	41	41
Limousin	311 660	1,31	13	13
Lorraine	908 731	3,82	38	38
Midi Pyrénées	1 070 768	4,50	45	45
Nord Pas de Calais	1 491 153	6,26	63	63
Pays de la Loire	1 292 594	5,43	54	54
Picardie	701 031	2,94	29	29
Poitou Charantes	687 280	2,89	29	29
Provence Alpes C.A	1 896 150	7,96	80	80
Rhône Alpes	2 273 841	9,55	96	96
Total	23 808 072	**100**	**1 000**	**1 000**

TAB. 3.1 – Répartition des ménages dans les 22 régions de France métropolitaine selon le recensement de 2006 (source INSEE). Les deux dernières colonnes représentent le nombre de personnes à sonder par région pour les deux populations.

Au final une base de numéros en liste annuaire, en liste rouge et de mobiles only est constituée. Elle est basée sur les proportions ou quotas relatifs aux ménages des 22 régions administratives de France métropolitaine (tableau 3.1). Ces proportions de cadrage permettent de fixer le nombre d'individus à recruter par région sur la base de 1 000 personnes par groupe : 1 000 enfants (de moins de 15 ans) et 1 000 adultes (de 15 ans et plus).

3.2.3.2 Recrutement des personnes

La participation à cette étude est basée sur le principe du volontariat. Aucune rémunération n'a été proposée à un individu pour réaliser les mesures. Un dédommagement, quel que soit sa nature, aurait pu être considéré comme un biais de recrutement.

L'ensemble des numéros constituant la base initiale est chargé dans un automate d'appel qui les compose de manière totalement aléatoire. Une fois le contact établi, un télé enquêteur (ou recruteur) se charge d'expliquer le but de l'étude à l'adulte de référence (ou chef de ménage) afin de proposer la participation d'un membre du foyer pour sa réalisation. Un screener de recrutement (questions filtres permettant de qualifier le contact) a été élaboré par MV2 Conseil. Si le contact répond positivement aux filtres du screener et accepte le principe de l'étude, le recruteur relève les dates de naissance de l'ensemble des membres du ménage. Celui dont la date de naissance est la plus proche de la date de prise de contact, est la personne *élue* pour porter l'appareil de mesure des CM. De ce fait, seul le hasard désigne la personne incluse dans le recueil des mesures de CM.
Si la personne élue refuse ou est dans l'incapacité de réaliser le test de 24 heures, le contact est considéré comme définitivement perdu.

3.3 Mesure des champs magnétiques

3.3.1 Choix de l'appareil de mesure

Pour réaliser les mesures, nous avons opté pour l'EMDEX II (Enertech, USA)(figure 3.1) du fait de sa performance par rapport à l'EMDEX Lite. Ce choix est guidé par des expériences réalisées au département d'électromagnétisme de SUPELEC sur ces deux types d'EMDEX [28]. L'EMDEX II mesure deux composantes de CM (large bande et harmonique) alors que l'EMDEX Lite ne mesure que la large bande. Ces deux composantes permettent d'identifier certaines sources en fonction des champs magnétiques enregistrés. Par exemple, une ligne à haute tension génère du 50 Hz (sans harmoniques) tandis que un radio-réveil génère environ 1/3 d'harmonique par rapport la composante 50 Hz. De plus l'EMDEX Lite est très sensible face aux ondes générées par les téléphones portables s'il n'est pas dans sa housse de protection. Le tableau 3.2 donne les caractéristiques de l'EMDEX II. Au total, 65 EMDEX II sont utilisés pour réaliser l'ensemble des mesures.

FIG. 3.1 – Photo d'un EMDEX II.

Fonction	**Spécificité**
Appareil	Système de mesure des CM
Enregistrement	Oui
Données	Mesures en temps réel
Gamme	0,01 à 300 μT
Résolution	0,01 μT
Précision	$\pm 1\%$
Mémoire interne	156 Kb ou 512 Kb
Fréquence	Large bande : 40-800 Hz Harmonique : 100-800 Hz
Période minimale d'échantillonnage	1,5 secondes
Dimensions	16,8 x 6,6 x 3,8 cm
Poids	341 grammes

TAB. 3.2 – Principales caractéristiques de l'EMDEX II.

3.3.2 Étalonnage et vérification des EMDEX

Après réception, les EMDEX ont été tous étalonnés dans un laboratoire du département d'électromagnétisme de SUPELEC. Dans ce laboratoire, le CM résiduel est très faible (entre 0,03 et 0,05 μT). Pour réaliser ces vérifications, deux systèmes semi-automatiques d'étalonnage ont été réalisés (figure 3.2). Ils sont formés chacun de deux ensembles de bobines de Helmholtz du commerce (Leybold), un générateur à basse fréquence programmable, un ampèremètre. Au centre de chaque système, le champ magnétique B en μT est donné par (3.1) où N_0 est le nombre de spires, R le rayon des bobines et la distance entre les centres des deux bobines de chaque système (en mètre), I l'intensité du courant (en Ampère) généré et $\mu_0 = 4\pi \times 10^{-7}$. L'écart entre la valeur affichée par l'EMDEX et celle donnée avec la formule (3.1) est mesurée. Cet étalonnage a été réalisé au début de l'étude, à la fin de l'étude et entre chaque campagne annuelle.

$$B = \mu_0 \frac{0,716 \times N_0 \times I}{R}. \tag{3.1}$$

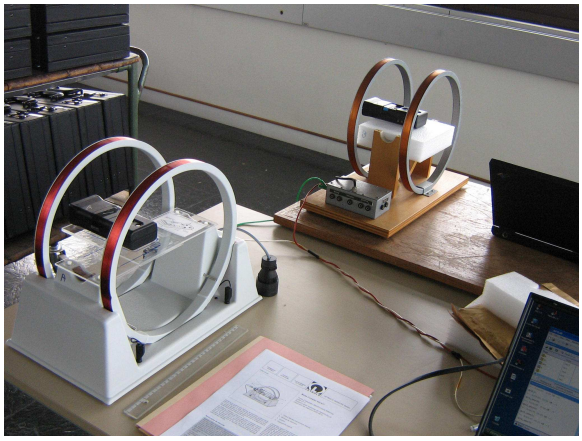

FIG. 3.2 – Système d'étalonnage des EMDEX.

Les caractéristiques des bobines utilisées sont $N_0 = 130$, la distance séparant les centres des deux bobines est $R = 115$ mm et l'intensité maximale du courant à générer est $I_{max} = 2$ A.

La figure 3.3 donne les courbes d'étalonnage de 15 EMDEX et celle obtenue avec la formule (3.1) en fonction du courant généré. Elle montre que les 16 droites sont totalement confondues. Des écarts négligeables apparaissent lorsque l'intensité du courant généré est faible. On en conclut que les écarts entre les valeurs observées et théoriques sont négligeables.

Au cours des trois années d'étude, aucun écart n'a été observé sur aucun des 65 EMDEX II utilisés.

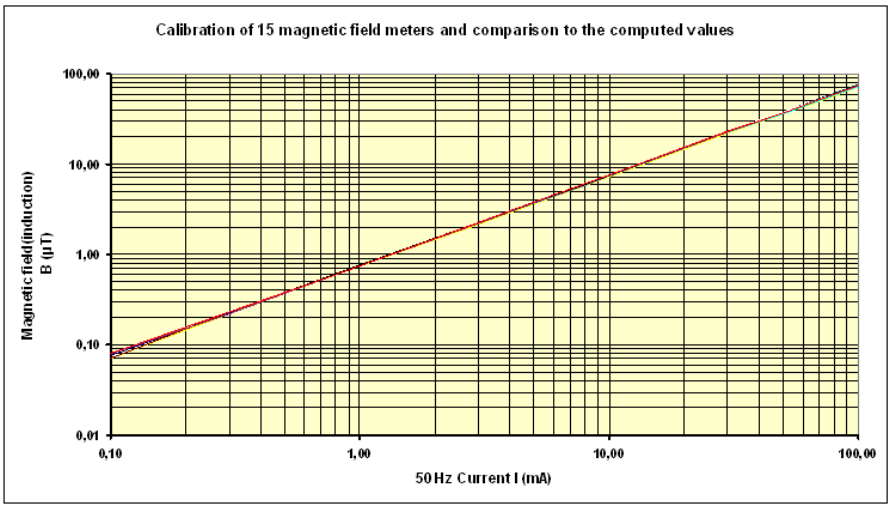

FIG. 3.3 – Comparaison des courbes d'étalonnage de 15 EMDEX.

Un autre système a été conçu pour vérifier si l'EMDEX est opérationnel après chaque expérience de 24 heures et avant de l'utiliser pour d'autres mesures. La figure 3.4 donne une photo de cette valise de test et le schéma du système. Le modèle de ce système est composé de solénoïdes carrés de 25 cm de côté avec 225 spires et 4 positions de courant. Le système permet de comparer la valeur théorique et celle affichée sur l'écran de l'EMDEX (et ce pour les trois positions de l'EMDEX, vérifiant que les trois bobines de l'EMDEX orientées selon les trois coordonnées cartésiennes sont en bon état). Il est conçu et recopié avec l'autorisation du propriétaire, l'australien Thanh Dovan (SWPNet). Pour avoir la traçabilité de l'EMDEX, l'enquêteur note sur le questionnaire les valeurs affichées sur l'écran de l'EMDEX pour chaque axe et ceci pour différentes valeurs de l'intensité du courant (voir annexe « Questionnaire »).

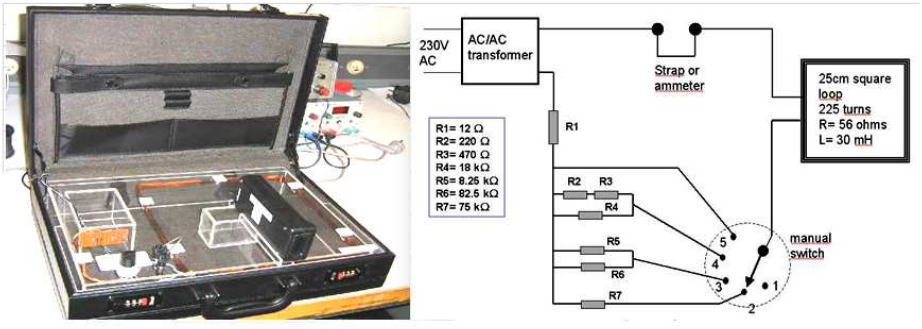

FIG. 3.4 – Photo et système de vérification des EMDEX sur le terrain.

3.3.3 Protocole de mesures

La personne volontaire porte un EMDEX II mesurant et enregistrant toutes les 3 secondes les CM auxquels elle est exposée. Le résultat enregistré est la norme du CM B sur les trois axes (3.2) avec deux composantes : la composante large bande (broad band resultant) et la composante harmonique (harmonic resultant).

$$B = \sqrt{B_x^2 + B_y^2 + B_z^2}. \tag{3.2}$$

Il a été demandé à chaque personne de porter l'EMDEX sur elle pendant la journée (une sacoche était fournie) et de poser près d'elle la nuit mais à 50 cm au moins de tout appareil électrique.

La figure 3.5 est un exemple d'une série de CM enregistrés par un volontaire. Les moyennes arithmétique et géométrique calculées sur la composante large bande sont respectivement de 0,08 et 0,05 μT. Cette figure présente des paquets d'observations qui peuvent correspondre à diverses activités. Pour identifier ces activités, on utilise le planning ou emploi du temps de la personne concernée.

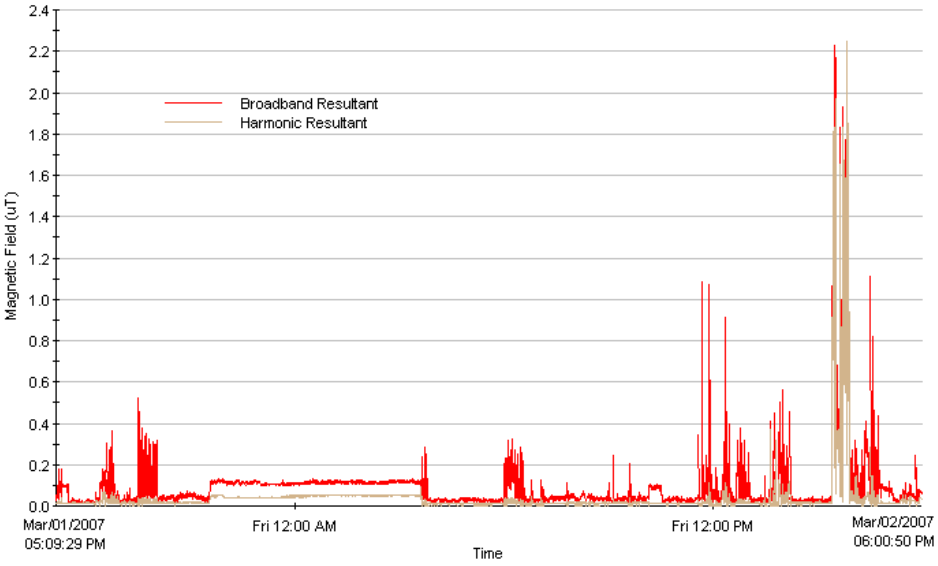

FIG. 3.5 – Exemple d'une série de CM enregistrés par un volontaire.

3.4 Informations relatives aux volontaires et à leurs activités

Pour pouvoir identifier les sources et mieux caractériser les niveaux d'exposition, on doit connaître les facteurs locaux liés, dans le temps, aux activités et aux lieux fréquentés le jour de l'enquête. D'autres informations liées quotidiennement à la personne sondée sont aussi nécessaires (environnement électrique du foyer et mode de vie).

3.4.1 Lieux fréquentés et activités menées pendant l'enregistrement

Un emploi du temps de la forme d'un livret est rédigé (voir annexe "Emploi du temps"). Il contient, entre autre, des instructions concernant le port de l'EMDEX, un exemple d'un emploi du temps rempli et un emploi du temps vierge à remplir par le volontaire. Une fois l'EMDEX mis en marche, le volontaire remplit au fur et à mesure l'emploi du temps (plages horaires, activités et lieux où elles sont effectuées et appareils électriques utilisés). Avec cet emploi du temps, on peut découper chaque série de CM selon les types d'activité, calculer le temps passé dans les différentes activités et la moyenne des CM associée.

Le tableau 3.3 est l'emploi du temps rempli par la personne ayant enregistré la série de la figure 3.5. À l'aide de ce tableau, on peut associer chaque paquet d'observations à son activité. Ce qui permet de dire, par exemple, que les valeurs les plus élevées sont observées quant la personne faisait ses courses.

Pour avoir plus de précisions sur les heures de début et de fin de chaque activité, nous avons introduit les deux dernières colonnes lors de l'exploitation des données. Elles correspondent respectivement aux numéros d'observations de début et de fin de chaque activité. À partir de ces numéros, on peut calculer le temps réel passé dans chacune des activités du planning des volontaires.

Début	Fin	Activité	Lieu	Appareils	N° d'observation	
17h10	19h10	Ordinateur	Domicile	Ordinateur	1	2411
19h10	22h00	Télévision	Domicile	Télévision	2412	5811
22h00	03h45	Sommeil	Chambre		5812	12711
03h45	09h00	Ordinateur	Domicile	Ordinateur	12712	19011
09h00	10h00	Déjeuner	Domicile		19012	20211
10h00	11h45	Ordinateur	Domicile	Ordinateur	20212	22311
11h45	14h20	Courses	Magasin		22312	25411
14h20	15h25	Déjeuner	Domicile		25412	26711
15h25	16h45	Sortie	Rue		26712	28311
16h45	18h00	Ordinateur	Domicile	Ordinateur	28312	29828

TAB. 3.3 – Emploi du temps de la personne ayant enregistré les CM représentés dans la figure 3.5.

3.4.2 Mode de vie et environnement électrique du foyer

Un questionnaire a été rédigé afin que l'enquêteur, avec l'aide du volontaire renseigne l'environnement électrique général du foyer ainsi que des informations relatives à la personne sondée (voir annexe « Questionnaire »). Ces dernières sont principalement le mode de vie du volontaire (l'âge, le sexe, le code professionnel, le type d'habitation, le nombre de personnes vivant dans le foyer, l'année de construction de l'habitation, le nombre d'étages, etc.). Pour les équipements électriques du foyer, on s'est orienté vers les équipements utilisés pendant la réalisation des mesures (type et énergie de chauffage du logement (éventuellement l'appareil utilisé pour le chauffage électrique), énergie et mode de chauffage de l'eau, etc.).
Comme la consommation en énergie électrique est plus élevée en période hivernale par rapport aux autres saisons, on peut penser que les CM correspondant à cette période sont plus élevés par rapport aux autres saisons. C'est pour cette raison que les mesures sont réalisées en hiver. Pour des questions de budget (humain et financier) mais surtout à cause des difficultés de recrutement, les mesures ont été réalisées en trois phases (février-avril 2007, octobre 2007-avril 2008 et octobre 2008-janvier 2009).

À noter que l'élaboration de l'emploi du temps et du questionnaire n'a pas été finalisée qu'après la fin de la première phase (après que 500 personnes aient été sondées). Ceci a permis d'améliorer l'emploi du temps et le questionnaire en demandant plus d'informations dans les phases qui ont suivi.

Les radio-réveils sont problématiques car ils peuvent générer à leur contact des champs magnétiques de plusieurs dizaines de micro Tesla (voir 4.2.1.2).

Rappelons qu'il était demandé aux volontaires de poser l'EMDEX à au moins 50 cm de tout appareil électrique pendant la nuit. C'est pourquoi des informations relatives à la chambre à coucher sont demandées telles que :
- Avez-vous un réveil électrique proche du lit ?
- Si oui, est-ce que l'EMDEX était à moins de 50 cm du réveil ?
Ces mêmes questions sont posées pour tout autre appareil électrique sous tension pendant que la personne dormait.

Pour identifier l'existence ou non d'ouvrages électriques (visibles ou non) à proximité du foyer, l'enquêteur note aussi les coordonnées GPS (Global Positioning System) observées à la porte d'entrée du domicile de la personne sondée. À partir de ces coordonnées, RTE et ERDF nous renseignent sur la présence ou non de lignes à haute, moyenne ou à basse tension (aériennes ou souterraines), de postes électriques ou encore de réseaux ferrés électrifiés à proximité du domicile. Le buffer utilisé autour du domicile est de 200 m pour les lignes à 400 kV, de 120 m pour les lignes à 225 kV, de 100 m pour les lignes à 150 kV, de 70 m pour celles à 90 ou 63 kV, de 20 m pour les lignes à moyenne tension de 20 kV et les lignes à basse tension de 380 V et 200 m pour les reseaux ferrés électrifiés. Pour les câbles souterrains, la distance maximale est aussi fixée à 20 m. Par exemple RTE trace un buffer de 200 m de part et d'autre de toutes les lignes aériennes 400 kV et regarde, dans la liste de tous les sujets, ceux dont les coordonnées GPS sont dans le buffer. La largeur du buffer a été calculée par RTE à partir du courant moyen annuel et de la géométrie des ouvrages générant le plus de champ magnétique, pour obtenir un champ de 0,1 μT en limite du buffer.

Cette recherche a permis d'identifier 9 personnes qui ont leurs foyers proches des lignes à 400 kV, 13 à côté des lignes aériennes à 225 kV, 22 à côté des lignes aériennes de 63, 90 ou 150 kV. Pour les câbles souterrains, 17 personnes habitent à côté des liaison de 225 kV et 20 personnes à côté des liaisons de 63, 90 ou 150 kV. On a identifié aussi 162 personnes (81 enfants et 81 adultes) qui habitent à côté des réseaux ferrés électrifiés. Ces données ont été fournies par RTE en juin 2009. Le travail de recherche des lignes à moyenne et basse tension et des postes pouvant se trouver près des foyers n'est pas encore finalisé. Cette base de données sera disponible vers mi 2010. Nous devons souligner que la fréquence des réseaux ferrés électrifiés n'est pas identifiée (alternatif ou continu) car les lignes sont trouvées à partir de cartes IGN (Institut Géographique National). C'est également pour cela qu'un buffer large (200 m) a été pris pour les réseaux ferrés électrifiés.

Pour la suite de l'étude présentée ici, nous étudierons seulement les réseaux 63 kV à 400 kV et les réseaux ferrés électrifiés. Avec ces informations, on peut étudier le lien entre les CM moyens observés au domicile et la présence ou non de ces ouvrages à proximité du foyer.

3.5 Base de données obtenue

Pendant la première phase, l'institut de sondage n'a pas pu recruter le nombre d'enfants escompté. Pour espérer avoir le nombre d'enfants et d'adultes en temps voulu (3 ans), le protocole de recrutement a été légèrement modifié en privilégiant le recrutement des enfants. Une fois le numéro de téléphone tiré au sort et une fois l'accord du chef de ménage pour la participation d'un membre de la famille obtenu, l'enquêteur demande si le ménage a un enfant. Si tel est le cas, il propose que l'EMDEX soit porté par l'enfant. Pour rassurer la famille, il propose aussi qu'un des parents porte un autre EMDEX. Comme les deux études (enfants et adultes) sont séparées, il n'y a pas de risque de corrélation arbitraire entre les CM de deux personnes du même ménage.

3.5.1 Analyse des numéros exploités

Pour recruter les volontaires, MV2 Conseil a dû construire une base de sondage de 95 362 numéros de téléphone (73 430 numéros sur annuaire, 14 304 numéros en liste rouge et 7 628 numéros de portable « only ») tout en respectant les proportions relatives aux régions (tableau 3.1). Ces numéros sont ensuite chargés dans un automate d'appel qui se charge de les composer de manière aléatoire.
La figure 3.6 représente la répartition des numéros exploités par l'automate d'appel. Au total, 44 437 appels ont abouti (soit $46,64\%$ des numéros chargés dans l'automate) et 3 047 personnes ont dans un premier temps accepté le principe de l'étude. Le taux de réussite ou d'acceptation du principe de l'étude est donc de $6,86\%$. Le temps moyen passé au téléphone pour le recrutement d'un volontaire est de 70 minutes. Le nombre de personnes effectivement sondées est de 2 148. Cela correspond à $4,83\%$ des appels décrochés et à $2,25\%$ des numéros composés par l'automate.

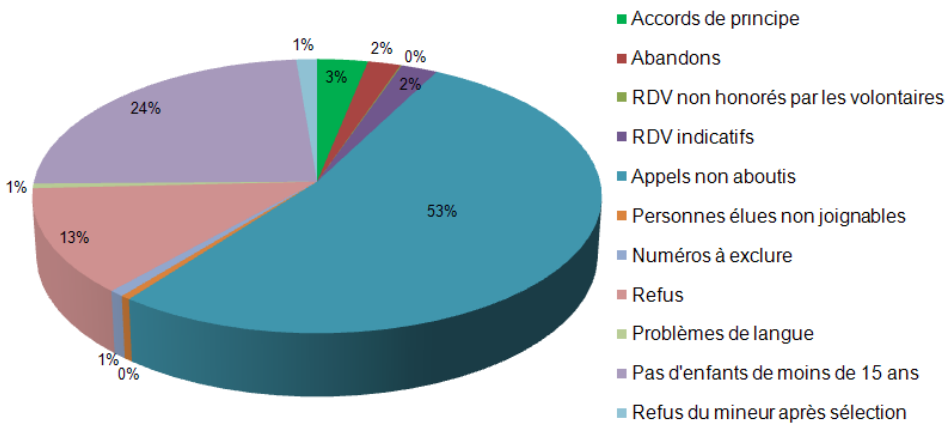

FIG. 3.6 – Répartition des numéros exploités par MV2 Conseil.

Cette enquête a permis de se rendre compte de la difficulté à recruter des sujets pour réaliser des mesures personnelles de CM sur 24 heures. Globalement, 53% des appels réalisés n'ont pas abouti (répondeur, numéro de fax, ...) et seulement 3% des personnes contactées ont accepté le principe de l'étude (figure 3.6). Il faut souligner que cette étude demande une participation importante des volontaires. Pour avoir 2 148 volontaires, l'institut de sondage a dû composer 95 362 numéros. Les principales difficultés rencontrées sont le recrutement des enfants. En effet il était indispensable d'obtenir l'accord des parents. Ces derniers acceptaient souvent pour eux-mêmes, mais refusaient que leur enfant fasse l'objet d'une recherche scientifique ou par peur que l'appareil de mesure soit cassé par l'enfant ou émette des ondes qui seraient nocives pour leur enfant, ceci, malgré une lettre d'information rédigée à cet égard par la DGS (voir annexe « lettre de la Direction Générale de la Santé »).

3.5.2 Validation de la base de données

Pour réaliser les mesures sur les 2 148 volontaires, SUPELEC a mis à disposition 65 EMDEX II dont 40 ont été loués à HPA (Health Protection Agency, Grande Bretagne, http ://www.hpa.org.uk).
Une fois les mesures réalisées, les emplois du temps et les questionnaires remplis, MV2 Conseil procède à des contrôles de validité de toutes ces informations. Plusieurs raisons ont conduit à éliminer 58 enregistrements chez les enfants et 40 chez les adultes (démontage de l'EMDEX, EMDEX non porté toute la période de mesure, emplois du temps non remplis, etc.) soit un taux de rebut de $5,47\%$ et $3,68\%$ (respectivement pour les enfants et les adultes). Le tableau 3.4 donne la répartition des mesures validées par MV2 Conseil dans les différentes régions. Ce tableau montre quelques légères différences

entre le nombre de mesures prévues et effectuées chez les enfants mais elles ne vont pas au-delà d'une unité (Franche Comté, Haute Normandie, Limousin). Chez les adultes ces différences sont importantes (216 au lieu de 189 en Île-de-France ou encore 111 au lieu de 96 en Rhône-Alpes). Ces deux régions sont les plus peuplées de France, les plus industrialisées et aussi les plus équipées en matière d'infrastructures de transports (RER, Métro, etc). Ces différences peuvent entraîner une surestimation de l'exposition moyenne de la population adulte.

	Nombre d'enfants		Nombre d'adultes	
Région	Prévu	Réalisé	Prévu	Réalisé
Alsace	29	29	29	29
Aquitaine	51	52	51	52
Auvergne	23	24	23	24
Basse Normandie	24	24	24	25
Bourgogne	28	28	28	28
Bretagne	51	51	51	51
Centre	42	42	42	42
Champagne Ardenne	23	23	23	23
Corse	4	4	4	4
Franche Comté	19	18	19	19
Haute Normandie	29	28	29	29
Île de France	189	189	189	216
Languedoc Roussillon	41	41	41	40
Limousin	13	12	13	12
Lorraine	38	38	38	40
Midi Pyrénées	45	45	45	45
Nord Pas de Calais	63	63	63	63
Pays de la Loire	54	56	54	55
Picardie	29	29	29	29
Poitou Charantes	29	30	29	29
Provence Alpes C.A	80	80	80	82
Rhône Alpes	96	96	96	111
Total	1 000	**1 002**	1 000	**1 048**

TAB. 3.4 – Répartition des mesures prévues et réalisées dans les 22 régions de France métropolitaine (et Corse).

La localisation des foyers de résidence des volontaires ayant réalisé les mesures sur le territoire national montre que 11 départements sur 96 ne sont pas représentés. Les numéros de ces départements sont marqués en rouge dans la figure 3.7. Les régions auxquelles appartiennent ces départements sont l'Auvergne (1 département sur 4), la Champagne Ardenne (1 sur 4),

le Centre (1 sur 6), la Corse (1 sur 2), le Languedoc Roussillon (1 sur 5), le Limousin (1 sur 3), la Provence Alpes Cote d'Azur (2 sur 6) et le Midi Pyrénées (3 sur 8). Ceci s'explique par le fait qu'aucun quota n'a été imposé sur les départements lors de la sélection des individus (quotas par régions seulement). Cela dit, ne pas avoir d'individus sur 11 départements peut paraître anormal et donc révéler un biais de sélection. Pour vérifier laquelle de ces hypothèses est la plus probable, des tests statistiques ont été réalisés. Ils sont basés sur le nombre de ménages dans les départements en question.

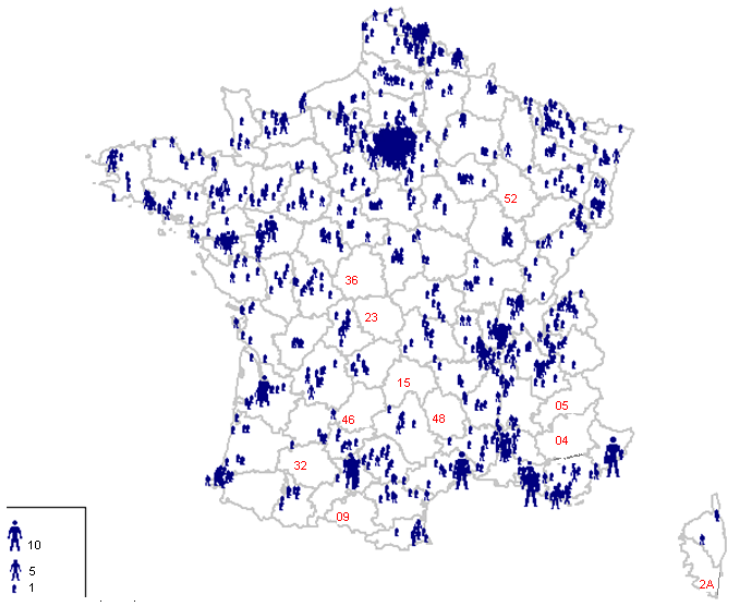

FIG. 3.7 – Localisation des volontaires sur la carte de la France.

À partir des proportions des ménages des régions d'où sont issus ces départements et de la population de chacun d'eux, nous avons estimé le nombre de ménages que compte chacun de ces départements (recensement de 2006, source INSEE). Soient $p_{E,k}$ et $p_{A,k}$ la probabilité qu'un enfant et un adulte respectivement soit sélectionné dans la région d'où est issu le département k, $k \in D_{NRP} = \{04, 05, 09, 15, 23, 32, 36, 46, 48, 52, 2A\}$. Connaissant le nombre total de ménages prévu et réalisé dans chaque région (tableau 3.1), il est possible de calculer la probabilité de sélectionner un ménage dans chacune des régions (p_E pour les enfants et p_A pour les adultes) et tester $p_{E,k} = p_E$ contre $p_{E,k} < p_E$, ou $p_{A,k} = p_A$ contre $p_{A,k} < p_A$.

Pour un département numéro $k \in D_{NRP}$ ayant n_k ménages, le nombre d'enfants (resp. d'adultes) sélectionnés suit une loi binomiale $\mathfrak{B}(n_k, p_E)$,

$\mathfrak{B}(n_k, p_A)$ sous l'hypothèse d'égalité (ou d'homogénéité). Pour un seuil $\alpha = 5\%$, le test rejette cette hypothèse si le nombre d'individus sélectionnés dans le département numéro $k \in D_{NRP}$ est inférieur au quantile à α d'une loi binomiale $\mathfrak{B}(n_k, p_E)$ pour les enfants et $\mathfrak{B}(n_k, p_A)$ pour les adultes. Une condition équivalente au rejet de l'hypothèse de l'homogénéité est le fait d'avoir une p-value inférieure à α. Pour les enfants, la p-value du test est la probabilité qu'une variable aléatoire suivant une loi $\mathfrak{B}(n_k, p_E)$ soit inférieure au nombre d'enfants observé dans le département numéro k.

Les résultats de ces tests sont représentés dans le tableau 3.5. Ils montrent que les probabilités $p_{E,k}$, $k = 32, 36, 52$ sont inférieures à p_E et que les $p_{A,k}$, $k = 32, 36, 46, 52$ sont aussi inférieures à p_A, avec un risque de se tromper de $\alpha = 5\%$. Si tel n'était pas le cas, on devrait observer au moins un individu pour chaque type de population dans ces départements.

Département		p-value en %		
Numéro	Nb de ménages	Enfants	Adultes	$\frac{i}{M}\alpha$ en %
36	96 226	1,8 (1 enfant)	1,8 (1 adulte)	0,5
52	77 931	3,6 (1 enfant)	3,6 (1 adulte)	0,9
32	72 529	4,7 (1 enfant)	4,7 (1 adulte)	1,4
46	70 416	5,2	5,2	1,8
15	64 999	6,1	6,1	2,3
04	63 872	6,8	6,3	2,7
09	58 305	8,6	8,6	3,2
05	54 330	10,1	9,5	3,6
23	55 866	11,6	11,6	4,1
2A	49 405	15,6	15,6	4,5
48	31 140	26,8	27,6	5,0

TAB. 3.5 – Calcul des probabilités des p-values des tests.

Ces résultats montrent que l'hypothèse d'homogénéité est rejetée uniquement sur trois départements. Toutefois les p-values ne sont pas très faibles (si on avait une seule personne par type de population dans chacun des départements 36, 52 et 32, les sélections seraient considérées homogènes).

Pour conclure, nous appliquons le test multiple de Benjamini [29] qui porte sur l'espérance du rapport entre le nombre d'hypothèses rejetées à raison et le nombre total d'hypothèses rejetées (False Discovery Rate). Ce test consiste à ranger par ordre croissant les p-values observées et à comparer chacune d'elles à $\frac{i}{M}\alpha$ où i désigne le rang de la p-value, α le seuil du test ($\alpha = 5\%$) et M le nombre total de tests à réaliser (ici $M = 11$). Le test rejette toutes les hypothèses nulles telles que $p_{(i)} \leq \frac{i}{M}\alpha$ avec $p_{(i)}$ la p-value numéro i.

Les résultats montrent que les $p_{(i)}$ sont toutes supérieures à $\frac{i}{M}\alpha$ ce qui veut dire que, globalement, on ne rejette pas l'homogénéité des sélections. Le fait qu'on n'ait pas d'individus dans les départements en question peut être considéré comme dû au hasard pour les deux types de population.

Toutes les informations validées par MV2 Conseil (2 048 séries de CM ainsi que les emplois du temps et les questionnaires associés) ont été récupérés par SUPELEC. Elles sont saisies informatiquement. Cette saisie a permis d'affiner les plages horaires des emplois du temps en superposant chaque série de CM avec l'emploi du temps correspondant.

Des irrégularités sont ainsi identifiées sur les emplois du temps ou sur les séries elles mêmes (période de mesure inférieure à 24 heures, emploi du temps correspondant à un adulte alors que la personne sondée est un enfant, etc.). Elles ont conduit à éliminer les données de 16 individus. Au total, les données de 2 032 personnes (978 enfants et 1 054 adultes) sont validées.
Ces 2 032 personnes ont été sélectionnées sans quota. Cela nous permet de regarder leur profil par rapport à la population française.

3.5.3 Profils des volontaires

Pour avoir une idée de la représentativité de l'échantillon par rapport à la population française, nous comparons la répartition des volontaires selon des classes d'âges et selon le sexe par rapport à la population nationale.

3.5.3.1 Répartition des volontaires selon des classes d'âge

Les enfants de moins de 6 ans sont moins représentés par rapport à la population nationale (figure 3.8). Les différences entres les proportions de la population et nos données observées atteignent 10%. Ces constatations pourraient s'expliquer par la réticence des parents face à l'appareil de mesures dès lors que les mesures devraient être réalisées par enfants en bas âge. Pour rappel, l'enfant est censé porter l'EMDEX sur lui. Un enfant de 2-3 ans est difficilement contrôlable et de nombreux parents ont eu peur que l'EMDEX soit cassé et, dès qu'ils le pouvaient, orientaient vers un autre enfant plus âgé ou refusaient d'y participer. Ils avaient peur que l'enfant se blesse mais également pour le matériel. Ils ne voulaient pas que leur responsabilité puisse être engagée.

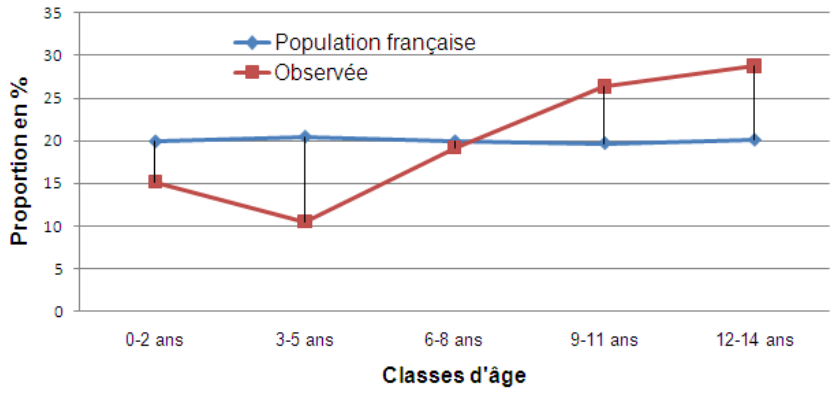

FIG. 3.8 – Comparaison des proportions de la population nationale et des proportions observées dans l'échantillon, selon des classes d'âge pour les enfants.

Pour la population adulte, les moins de 25 ans et les plus de 59 ans sont moins représentés (figure 3.9). Les différences les plus importantes sont observées sur les 38-44 ans où le nombre de personnes dans l'échantillon est nettement plus important que dans la population française.
Ces différences peuvent s'expliquer par la modification du protocole pour favoriser le recrutement des enfants. On dispose de 519 familles représentées par deux volontaires (un enfant et un adulte). Ceci induit une corrélation importante entre le recrutement des enfants et l'âge des adultes car ce sont majoritairement des parents de moins de 50 ans qui ont des enfants de moins de 15 ans.

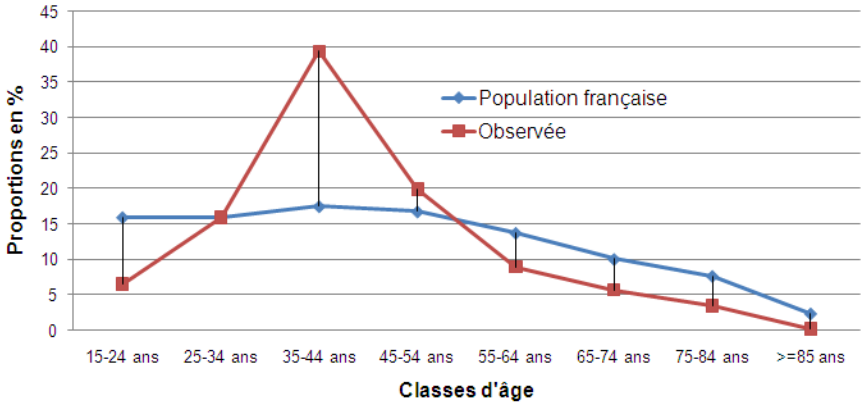

FIG. 3.9 – Comparaison des proportions de la population nationale et des proportions observées dans l'échantillon, selon des classes d'âge pour les adultes.

3.5.3.2 Répartition des volontaires selon le sexe

Comme dans le cas des classes d'âges, nous avons comparé de manière descriptive les proportions observées des volontaires par sexe, par rapport à la population française. Comme les seuls quotas imposés lors de la sélection des individus sont la répartition de la population sur les 22 régions, rien ne peut être pressenti sur la répartition homme femme.

La répartition des enfants de l'échantillon selon le sexe montre que 51% sont de sexe féminin contre 49% pour la population française métropolitaine (figure 3.10). Pour voir si cette différence est significative ou non, nous réalisons un test d'homogénéité. Le but est de voir si la proportion des enfants de sexe féminin est statistiquement la même que dans la population française, ou si elle est plus élevée. Sous l'hypothèse d'égalité, le nombre d'enfants de sexe féminin observé suit une loi binomiale $\mathfrak{B}(0,49,n_E)$ où n_E désigne le nombre total d'enfants dans l'échantillon. Le test rejette l'hypothèse nulle si la probabilité qu'une loi $\mathfrak{B}(0,49;n_E)$ dépasse le nombre d'enfants de sexe féminin observé est inférieure au seuil $\alpha = 5\%$. Cette p-value est de 0,064, on ne rejette pas l'homogénéité des deux populations. La différence observée sur les deux populations en termes de sexe n'est donc pas statistiquement significative.

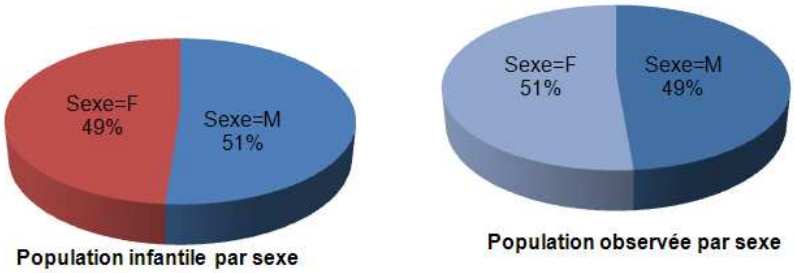

FIG. 3.10 – Comparaison des proportions de la population nationale et des proportions observées dans l'échantillon, selon le sexe pour les enfants.

Pour les adultes, la différence est plus importante. L'échantillon est composé de 64% de femmes contre 52% pour la population (figure 3.11). La p-value du test d'homogénéité est inférieure à 0,001. La différence entre les adultes de sexe féminin dans la population et dans l'échantillon est très significative. Cette différence peut être due au fait d'avoir facilité le recrutement des enfants en indiquant qu'un adulte du foyer pourrait aussi être inclus dans l'échantillon adulte. Même si les deux enquêtes sont totalement dissociées, on peut imaginer que les femmes acceptaient plus facilement ce principe que leurs compagnons. Toutefois, on ne prétendait pas avoir des

proportions similaires car la méthode utilisée ne tenait pas compte de ce facteur. Par ailleurs, la comparaison des expositions moyennes montre que le champ moyen observé pour les femmes est statistiquement équivalent à celui observé pour les hommes (voir 5.4). Le sexe n'est donc pas un facteur discriminant.

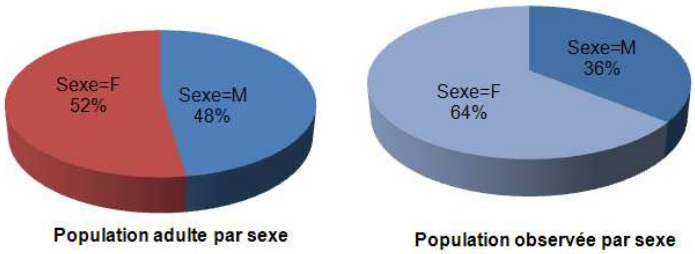

FIG. 3.11 – Comparaison des proportions de la population nationale et des proportions observées dans l'échantillon, selon le sexe pour les adultes.

3.6 Conclusion

Dans ce chapitre, nous avons présenté les différents choix considérés pour la construction de la base de données. La méthode utilisée pour la sélection des volontaires est la méthode de tirage aléatoire. L'institut de sondage a construit une base de sondage de 95 362 numéros de téléphone selon les proportions des ménages des 22 régions de France métropolitaine incluant des numéros sur annuaire, en liste rouge et de « portable only ». Ces numéros ont été composés de manière aléatoire par un automate d'appel. Dans l'ensemble, $6,86\%$ des numéros composés ont abouti, dans un premier temps, à un accord de principe pour participer à l'étude. Très souvent, des parents refusaient la réalisation des mesures dès lors que le choix se portait sur un enfant en bas âge. Cela a conduit à une modification du protocole en privilégiant le recrutement des enfants. Le nombre de volontaires ayant participé aux mesures des CM est de 2 148 (soit $2,25\%$ des numéros composés) et le temps moyen passé pour le recrutement d'un volontaire est de 70 minutes.

L'analyse de la pertinence de la base ainsi construite a permis de remarquer que 11 départements sur 96 ne sont pas représentés. Les tests d'homogénéité réalisés dans ces départements par rapport aux régions dont ils sont issus ont montré que ce fait n'induit pas un biais dans la base de données. Cela dit, les seuls quotas appliqués pour la sélection des individus restent les proportions des ménages dans les régions.

Une fois les informations recueillies, des contrôles de validité ont été

réalisés par MV2 Conseil et par SUPELEC qui se chargeait de la saisie. Le nombre de personnes dont les mesures ont été validées est de 2 032 (978 enfants et 1054 adultes). Ces contrôles ont permis aussi d'identifier 422 personnes qui ont posé l'EMDEX à côté d'un radio-réveil pendant la nuit, ce qui a induit un biais sur l'estimation de l'exposition comme nous le montrerons dans la section 4.2.

L'analyse des profils des volontaires a permis de conclure que la base de données dispose de moins d'enfants de moins de 6 et de plus d'enfants de 8 à 14 ans par rapport à la population nationale. Pour les adultes, l'échantillon a plus de personnes de plus de 34 ans et de moins de 50 ans, par comparaison à la population française. Pour le facteur sexe, les proportions dans l'échantillon et dans la population sont presque identiques pour les enfants alors que les femmes sont plus représentées que les hommes pour les adultes. Mais le sexe n'apparaît pas comme facteur discriminant en termes d'exposition.

Chapitre 4

Étude descriptive des expositions moyennes

4.1 Introduction

Dans ce chapitre, nous abordons l'un des objectifs de l'étude qui est l'estimation de l'exposition de la population. Cette notion fait référence aux expositions moyennes sur 24 heures alors que la notion de moyenne n'est pas bien précisée dans la majorité des études épidémiologiques (moyenne arithmétique (MA) ou moyenne géométrique (MG)). La définition même du terme « *exposition* » est difficile à définir du fait de la complexité de l'organisme humain (il comporte des zones électriquement excitables comme le cerveau et d'autres qui ne le sont pas comme les membres). Nous assimilons l'exposition à la mesure du champ magnétique en supposant que l'EMDEX était porté au plus près du corps. Pour étudier l'exposition selon les différents critères (enfants/ adultes, avoir son foyer à proximité des lignes de RTE/ à proximité des réseaux ferrés électrifiés/ loin de ces ouvrages, domicile/ extérieur, etc.), des tests de comparaison sont faits. On insistera sur les outils utilisés ainsi que sur les précautions prises pour la réalisation des tests. Pour clore ce chapitre, des estimations des expositions moyennes par activité sont calculées ainsi que les intervalles de confiance associés.

Nous ne pouvons pas justifier que les personnes ayant posé l'EMDEX à proximité des radio-réveils sont exposées ou non aux CM générés par ces derniers (voir section 4.2.1.2). Nous avons alors conduit l'étude de deux manières :

1. Considérer les CM enregistrés pendant toute la période de mesure.
   Dans ce cas, nous incluons des mesures liées aux radio-réveils. La principale conséquence est une possible sur-estimation de l'exposition.
2. Considérer les CM enregistrés en dehors de la période de sommeil.
   Pour chaque individu, on enleve les CM enregistrés pendant la période de sommeil en se basant sur son emploi du temps.
   Il est clair que, dans ce cas, nous éliminons en moyenne un tiers des observations de chaque série mais nous considérons que les CM analysés reflètent l'exposition des personnes.

4.2 Analyse descriptive des CM moyens

4.2.1 CM enregistrés sur 24 heures

4.2.1.1 Les expositions moyennes

Les hypothèses sur les risques pour la santé avancés par les scientifiques et l'Organisation Mondiale de la Santé (OMS) reposent sur une exposition supérieure à 0,4 μT en moyenne sur 24 heures chez l'enfant [2]. Du fait que la moyenne considérée n'est pas la même pour toutes les études, nous analysons les MA et les MG.

– **Les enfants**
Les expositions moyennes observées sur les enfants sont de 0,09 et 0,02 μT respectivement pour les MA et les MG. Au total, 30 enfants ont observé une MA supérieure à 0,4 μT soit $3,1\%$ des enfants. Parmi eux, deux ont observé une MG supérieure à cette valeur. La figure 4.1 donne la répartition des MA et des MG des enfants. Cette proportion est plus élevée par rapport à celle observée dans l'étude UKCCS (moins de $0,4\%$) [13] ou celle d'Ahlbom ($0,8\%$) [14]. C'est pourquoi nous avons cherché à expliquer ces expositions (voir 4.2.1.3)

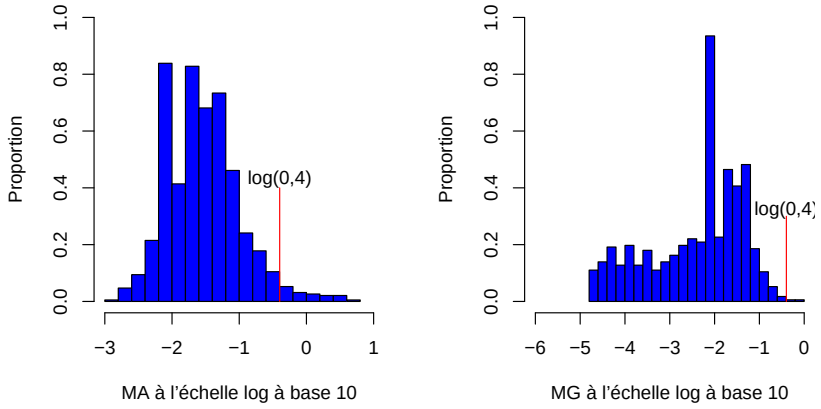

FIG. 4.1 – Histogramme des moyennes arithmétiques et géométriques observées par les enfants en μT.

Le calcul des quantiles des expositions moyennes (Tableau 4.1) montre que les valeurs médianes observées des moyennes sont respectivement de 0,03 et 0,01 μT pour MA et MG. Un pour cent d'entre eux, c'est-à-dire approximativement 10 enfants, ont observé une MA supérieure ou égale à 1,22 μT.

	25%	50%	75%	99%
Moyennes arithmétiques en μT	0,01	0,03	0,06	1,22
Moyennes géométriques en μT	0,00	0,01	0,02	0,19

TAB. 4.1 – Quelques quantiles des expositions moyennes des enfants.

– **Les adultes**
Pour les adultes, la répartition des moyennes arithmétiques et géométriques est donnée dans la figure 4.2. Les valeurs moyennes sont de 0,14 μT pour les MA et 0,03 μT pour les MG. Les valeurs médianes

sont de 0,05 μT pour les MA et 0,02 μT pour les MG. Un pour cent ou approximativement 11 adultes ont observé une MA supérieure ou égale à 1,54 μT (tableau 4.2).

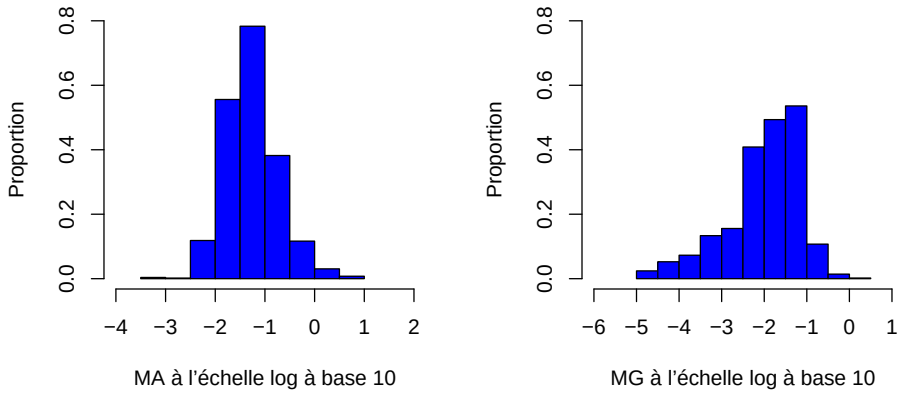

FIG. 4.2 – Histogrammes des moyennes arithmétiques et géométriques observées par les adultes en μT.

	25%	50%	75%	99%
Moyennes arithmétiques en μT	0,03	0,05	0,11	1,54
Moyennes géométriques en μT	0,00	0,02	0,04	0,26

TAB. 4.2 – Quelques quantiles des expositions moyennes des adultes.

4.2.1.2 Expositions de type radio-réveil

Lors de l'exploitation des données de la première vague, nous avons constaté l'existence de séries de CM liés à des radio-réveils. Les appareils électriques de type radio-réveil possèdent un petit transformateur associé à de l'électronique. Lorsque l'appareil de mesure est à proximité (quelques centimètres) d'un tel appareil, les CM enregistrés peuvent être importants (figure 4.4 avec des CM supérieurs à 15 μT pendant plusieurs heures). On comprend que de telles mesures conduisent à des CM moyens élevés et peuvent biaiser les résultats car ces derniers ne sont pas forcement représentatifs de l'exposition de la personne. Les CM 50 Hz générés par des appareils de type radio-réveils se traduisent par une composante harmonique d'environ 1/3 de la composante large bande. Nous avons associé toutes les expositions de ce type survenues pendant la nuit à des radio-réveils. Pour éviter de tels enregistrements dès la deuxième vague de mesures, il a été précisé par l'enquêteur et écrit dans l'emploi du temps de placer l'appareil de mesure sous le lit ou à

proximité de soi, en veillant à ce qu'il soit au moins à 50 cm de tout appareil électrique (radio, réveil, etc.).

Parmi les 2 032 volontaires dont les données sont retenues, 422 (147 enfants et 275 adultes) ont posé l'EMDEX à moins de 50 cm d'un radio-réveil ou d'un appareil électrique sous tension pendant toute la nuit ou la période de sommeil. Cela a été vérifié en superposant la plage horaire du signal et l'emploi du temps.

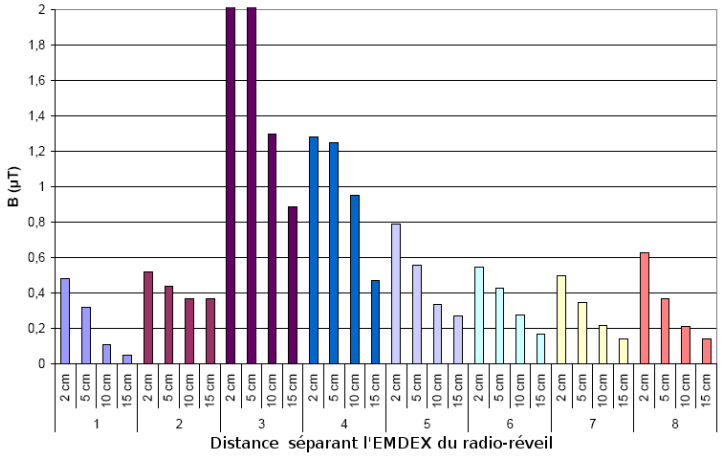

FIG. 4.3 – Zoom sur l'axe des ordonnées des mesures des CM enregistrés par un EMDEX en fonction de la distance séparant l'EMDEX du radio-réveil.

Le choix de proposer aux personnes de poser l'appareil de mesure à plus de 50 cm du radio-réveil vient d'une étude de mesures des CM générés par ce type d'appareil électrique en fonction de la distance qui les sépare. Cette étude a été réalisée au laboratoire des matériels électriques de EDF. Ces mesures sont réalisées avec un EMDEX et pour 8 radio-réveils chez les personnes travaillant au service des études médicales de EDF. La figure 4.3 donne les valeurs des CM enregistrés à 2 cm, 5 cm, 10 cm et 15 cm. Cette figure montre que le champ magnétique dépend du radio-réveil. Elle montre qu'un EMDEX posé au contact du radio-réveil mesurera un champ magnétique supérieur à 0,4 μT la nuit. Dans environ 5 cas sur 8, un EMDEX posé à 5 cm du réveil mesurera un champ magnétique supérieur à 0,4 μT.

Des EMDEX ont également été laissés toute une nuit contre des radio-réveils, afin de vérifier la stabilité du champ mesuré. Ces mesures réalisées avec le même EMDEX confirment que les niveaux de champ sont très va-

riables selon les modèles de réveils, mais également que dans certains cas le niveau de champ mesuré varie au cours de la nuit, parfois de 10 à 15%. La figure 4.4 donne un exemple d'une série de CM générés par un radio-réveil et enregistrés par L'EMDEX. Pour expliquer ces variations de champ magnétique, une expérimentation a été menée. Elle a été effectuée au domicile pour enregistrer simultanément la fluctuation de la tension du secteur qui est plus importante qu'au laboratoire (réseau BT étendu). Le but était de vérifier que les variations constatées sur de nombreux enregistrements du champ près d'un radio-réveil correspondent bien à la variation de sa tension d'alimentation.

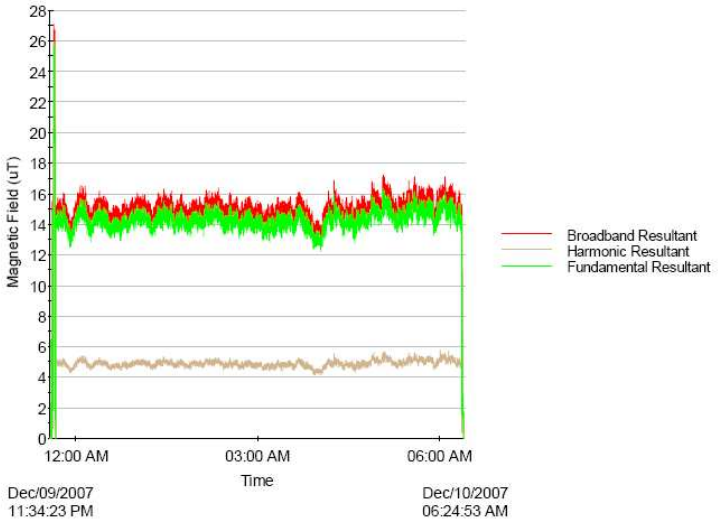

FIG. 4.4 – Exemple d'une série de CM générés par un radio-réveil.

Le radio-réveil Digicube SONY a été choisi car il génère un niveau de champ magnétique au contact parmi les plus élevés par rapport aux 8 autres de la figure 4.3. Il est ancien (> 20 ans) et le transformateur de son alimentation est intégré dans l'appareil. Ce n'est plus le cas des nouveaux radio-réveils où le transformateur est souvent déporté au niveau de la prise secteur ce qui diminue le champ généré au niveau du radio-réveil.

Dans cette expérience, la fluctuation de la tension du secteur est enregistrée avec un enregistreur thermique universel SEFRAM 8400 n° ME007086. La tension secteur est relevée au niveau d'un prolongateur équipé de 3 prises de courant qui sert à alimenter le radio-réveil et l'enregistreur (figure 4.5). L'enregistreur a été écarté volontairement du radio réveil pour qu'il ne perturbe pas la mesure de champ. Le calibre d'entrée de la voie 1 de mesure de la tension était réglé sur 250 V. La pleine échelle sur le papier correspondait à 0 V - 250 V (figure 4.6).

On observe que les deux courbes se superposent parfaitement. Les résultats sont identiques lorsque l'on fait varier la distance entre le radio-réveil et l'EMDEX. Les fluctuations observées sur les signaux de type radio-réveil correspondent donc aux fluctuations de la tension du secteur.

FIG. 4.5 – Mesure simultanée du CM (à gauche) et de la tension (à droite).

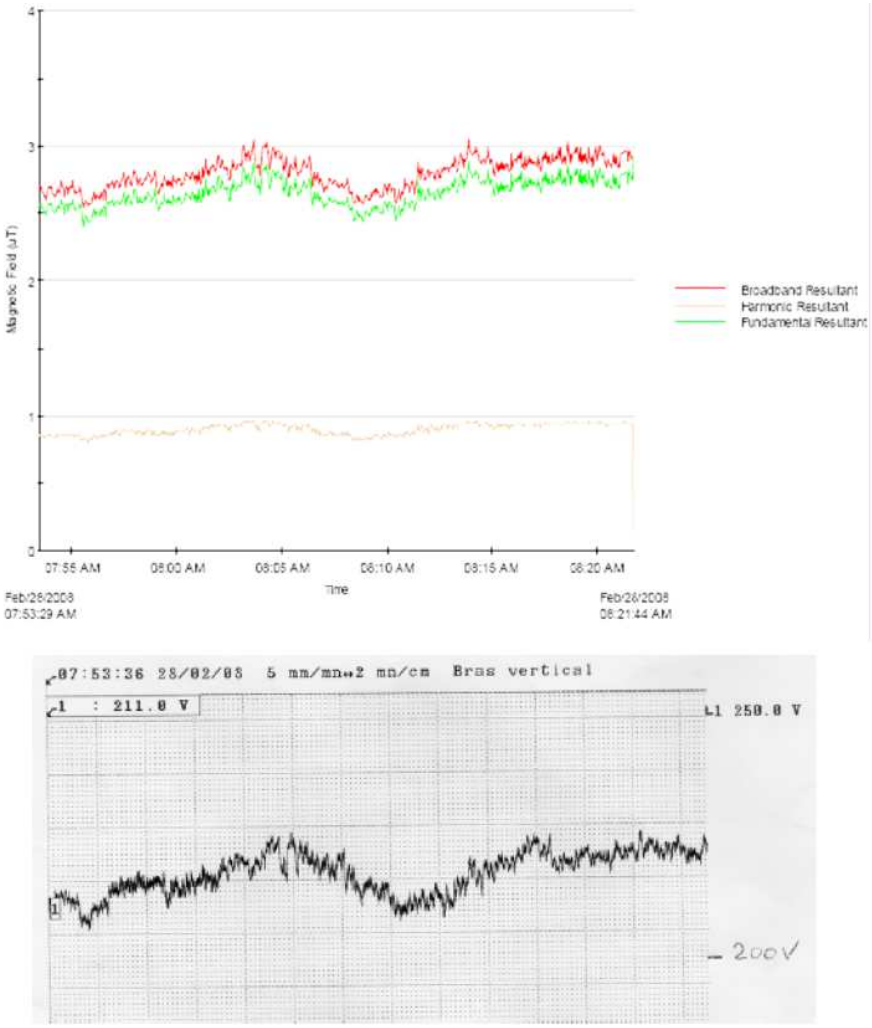

FIG. 4.6 – Enregistrement simultané du champ magnétique (en haut) et de la tension du secteur (en bas) au contact du radio-réveil.

Les résultats de cette expérience montrent que, pour ce radio-réveil, la décroissance du CM émis par le radio-réveil en fonction de la distance n'est plus mesurable à partir d'une distance de 50 cm du radio-réveil (figure 4.7).

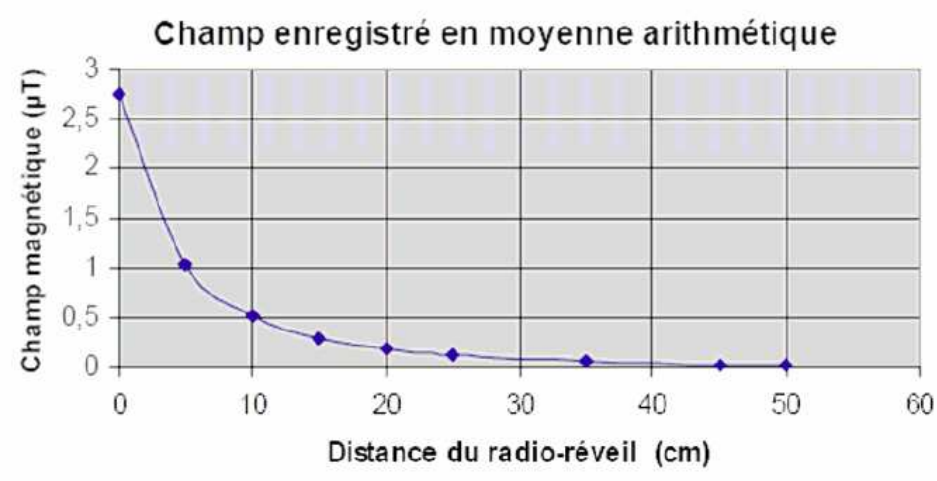

FIG. 4.7 – Décroissance du CM moyen généré par le radio-réveil avec la distance le séparant de l'EMDEX.

Suite aux interrogations sur le bien fondé de la distance de 50 cm à respecter entre l'EMDEX et le radio-réveil la nuit, MV2 Conseil, l'institut de sondage en charge du recueil des données, a gracieusement proposé de faire des mesures avec 2 EMDEX dans l'étude EXPERS. Elles sont réalisées lors du recueil des dernières mesures de l'étude EXPERS. La modification du protocole consistait en :

- un EMDEX qui mesure l'exposition personnelle comme prévu dans le protocole d'EXPERS,
- un EMDEX que le volontaire pose sur son oreiller pendant la journée (et où il veut pendant la nuit).

L'hypothèse est que s'il y a exposition au CM émis par un radio-réveil, le champ magnétique sur l'oreiller est plus faible que le champ magnétique la nuit avec un EMDEX posé près du radio-réveil. Au total, 38 mesures ont été réalisées selon ce protocole par 8 enquêteurs. Quatre ont été retirées pour erreur manifeste entre la courbe de champ magnétique et l'emploi du temps. Il en reste 34 qui se répartissent entre :

- 11 qui ne montrent rien (pas de radio-réveils, ou pas de différence manifeste entre oreiller et nuit)
- 5 avec un champ magnétique sur l'oreiller inférieur au champ magnétique la nuit
- 18 avec un champ magnétique sur l'oreiller supérieur au champ magnétique la nuit.

Ces résultats semblent montrer le contraire de l'hypothèse de départ. Le point étrange est que l'on a un certain nombre de mesures sans radio-réveil la nuit mais avec de fortes valeurs (jusqu'à 11 μT) sur l'oreiller, qui semblent

indiquer un EMDEX posé au contact d'un radio-réveil (avec des valeurs différentes avant et après la nuit : ce qui indique un EMDEX posé en un point légèrement different tout près d'une source ponctuelle).

Un autre point étrange est que ce phénomène concerne principalement 2 enquêteurs qui ont réalisé la moitié de ces mesures. MV2 a interrogé les enquêteurs qui ont dit avoir fait les mesures selon le protocole décrit ici (mais ce protocole n'a pas été donné par écrit aux volontaires).

Le comité technique d'EXPERS a conclu qu'il est impossible de tirer une information de ces mesures car la consigne n'a manifestement pas été bien comprise (peu importe que ce soit les volontaires ou les enquêteurs, c'est tout simplement non interprétable).

La mesure de champ sur les oreillers semble une bonne solution pour valider le critère de distance de 50 cm, mais il faut s'assurer que les mesures soient faites correctement. Il a donc été décidé de faire des mesures complémentaires autour de radio-réveils.

Ces mesures ont été réalisées par les membres du comité technique d'EXPERS, des collègues d'EDF R&D ou du SEM. Afin de s'assurer que les mesures sont faites correctement, un protocole détaillé leur a été fourni. Huit mesures ont ainsi été réalisées (figure 4.8). Il était demandé de faire des mesures depuis le radio-réveil jusqu'au milieu de l'oreiller, en mesurant les distance au radio-réveil.

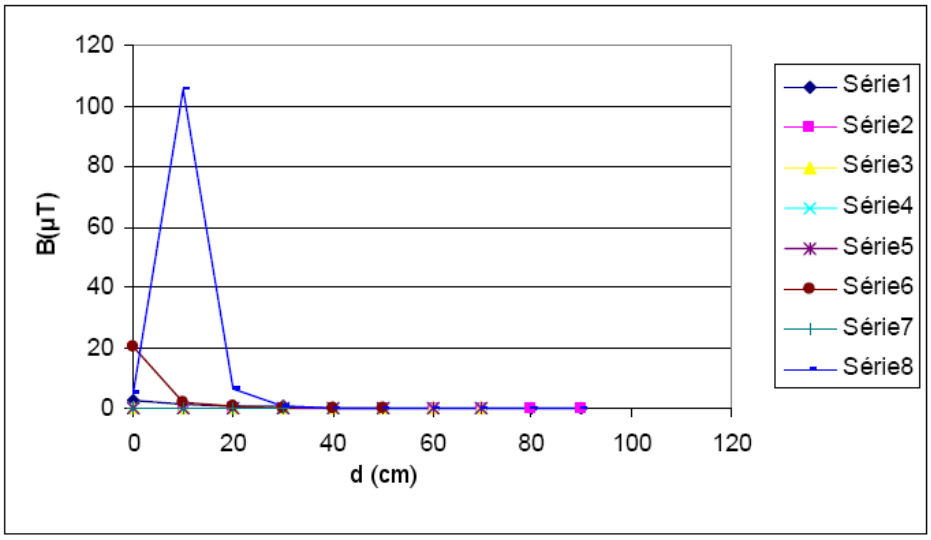

FIG. 4.8 – Champ magnétique émis par différents radio-réveils en fonction de la distance les séparant du centre de l'oreiller.

On observe que le champ magnétique est moins élevé si le transformateur est situé au niveau du bloc d'alimentation et non au niveau du radio-réveil. Le pic de la série 8 correspond au passage devant le bloc alimentation branché sur une prise en hauteur.

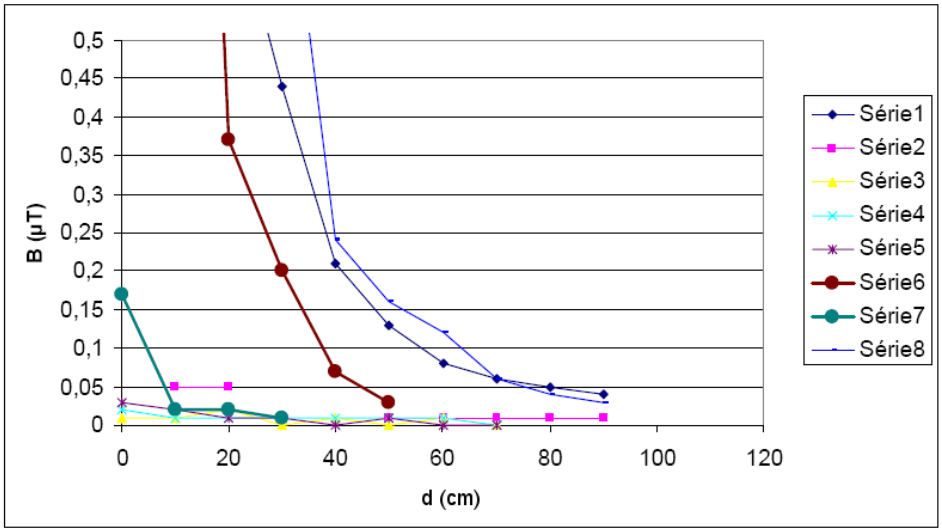

FIG. 4.9 – Champ magnétique mesuré du radio-réveil au centre de l'oreiller (zoom sur les valeurs des champs magnétiques inférieures 0,5 μT).

La figure 4.9 montre que dans 7 cas sur 8, le centre de l'oreiller est situé à une distance supérieure ou égale à 50 cm. Dans tous les cas, le champ magnétique au niveau de l'oreiller est dans le bruit de fond. La recommandation de poser l'EMDEX à au moins 50 cm du radio-réveil ne sous-estime pas l'exposition personnelle réelle.

Le choix de la distance de 50 cm a été validé par toutes ces séries d'expérience. Il a été donc demandé pour la suite de l'étude EXPERS de respecter une distance de 50 cm entre l'EMDEX et tout appareil électrique la nuit. Tous les volontaires n'ont pas respecté ce critère, et sur le total de l'étude EXPERS, 20% des mesures sont liées à des radio-réveils la nuit. La manière de traiter les signaux de type radio-réveil a été longuement discutée au comité technique d'EXPERS. Il est hors de question de supprimer les 20% de mesures concernées, car elles portent de l'information. Tout ce que nous pouvons dire est que l'exposition personnelle réelle est inférieure à l'exposition sur 24 heures telle que mesurée dans cette étude. Pour étudier les paramètres influençant l'exposition, et s'affranchir des radio-réveils, la solution adoptée est d'étudier aussi l'exposition hors période de sommeil.

4.2.1.3 Les sources identifiées

Pour les enfants, les principales sources associées aux expositions moyennes supérieures à 0,4 μT sont de différents types.
Parmi les 30 enfants ayant observé une moyenne arithmétique supérieure à 0,4 μT, 24 ont déclaré avoir posé l'EMDEX à moins de 50 cm d'un radio-réveil ou d'un appareil électrique sous tension pendant la nuit. En regardant les mesures de champ magnétique, on note la présence d'un radio-réveil pour 24 enfants. Deux d'entre eux habitent, l'un à proximité d'une ligne de 225 kV, l'autre, à côté d'une voie ferrée électrifiée (mais les mesures ne reflètent pas la présence de ces lignes).
Un enfant habite à côté d'un réseau ferré électrifié (et les mesures confirment que c'est la source de champ magnétique). La figure 4.10 donne les CM enregistrés par cet enfant. Elle montre que les CM enregistrés sont plus élevés sauf dans la période 01h30-04h30 du matin (le réseau n'est pas parcouru par du courant car il n'y a pas de trafic pendant cette période).
Quatre enfants ont comme sources des appareils électriques équipés de petits transformateurs.
Pour le dernier, la source semble être un réseau électrique, qui n'est ni un réseau à haute tension ni une voie ferrée (à confirmer avec l'étude de la proximité de réseaux basse et moyenne tension).

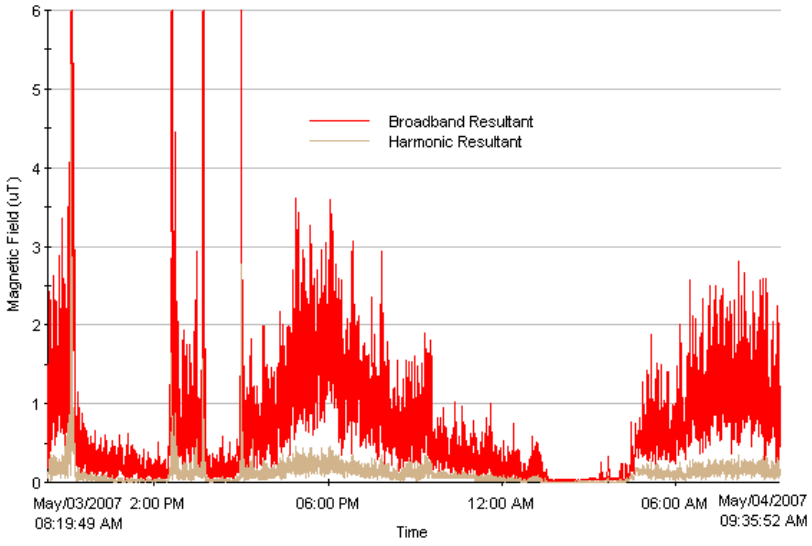

FIG. 4.10 – Zoom sur l'axe des ordonnées des CM enregistrés par un enfant ayant son établissement scolaire et son foyer de résidence à proximité d'un réseau ferré électrifié (MA= 0,66 μT et MG=0,37 μT).

Les enfants ayant observé une moyenne supérieure au quantile à 99%

(tableau 4.1) ont tous déclaré avoir posé l'EMDEX à côté du radio-réveil pendant le sommeil et les signaux reflètent bien cette information.

Parmi les 11 adultes ayant observé une MA supérieure au quantile à 99%, 9 ont indiqué que l'EMDEX était à côté du radio-réveil pendant la nuit, ce qui est confirmé par les mesures. Deux d'entre eux ont leur domicile à côté de réseaux ferrés électrifiés et un autre proche d'une ligne de 225 kV, mais les mesures ne reflètent la présence de ces lignes.
Pour une autre personne, la source principale est un appareil électrique équipé d'un petit transformateur utilisé dans son lieu de travail (école). Enfin, pour la dernière personne, il s'agit d'une source non identifiée liée à son domicile.

4.2.2 Champs magnétiques hors période de sommeil

1. **Les enfants :**
   En considérant les CM hors période de sommeil, les expositions moyennes des enfants sont de 0,05 μT pour les MA et 0,02 μT pour les MG. Les valeurs médianes observées sur ces deux moyennes sont respectivement de 0,03 et 0,01 μT. Onze enfants $(1,1\%)$ ont observé une moyenne arithmétique supérieure à 0,4 μT, 2 d'entre eux ont enregistré une MG supérieure à cette valeur. Sur ces 11 enfants, 8 ont observé une moyenne sur 24 heures supérieure à 0,4 μT, mais les sources d'exposition ne sont pas forcement les mêmes, puisqu'il n'y a pas *a priori* de réveil hors période de sommeil.
   On retrouve l'enfant avec le réseau ferré électrifié et celui avec le réseau électrique près de son domicile.
   Pour 7 enfants, les sources sont des appareils électriques équipés de petits transformateurs, utilisés au domicile (sauf pour un, à l'école).
   Pour un enfant, la source semble être un réseau électrique près de son école (à confirmer).
   Pour le dernier, des valeurs élevées ont été mesurées pendant un trajet de voiture. On suppose que l'EMDEX a été posé au contact d'un câble électrique pendant ce trajet. Ces mesures pourraient ne pas reflèter l'exposition de l'enfant.
2. **Les adultes :**
   Les moyennes arithmétique et géométrique observées chez les adultes sont respectivement de 0,10 et 0,03 μT. Les expositions médianes en termes de MA et MG sont respectivement de 0,05 et 0,02 μT. Le quantile à 99% des MA est de 0,83 μT, il est de 0,21 μT pour les MG.
   Parmi les 11 adultes ayant observé une MA supérieure au quantile à 99%, les sources sont principalement des appareils électriques équipés de petits transformateurs, utilisés au domicile ou au travail (8 adultes).
   On retrouve la personne avec la source inconnue au domicile.

Pour les 2 dernieres personnes, les sources sont des transportys ferroviaires. La figure 4.11 représente les CM enregistrés dans les transports ferroviaires par une de ces deux personnes, qui habite également près d'une voie ferrée. Pendant son trajet, cette personne était exposée à des CM de forte intensité dépassant même les 100 μT (seuil de l'ICNIRP et de recommandation de 1999). Les moyennes sur 24 heures observées par cette personne sont de 0,70 μT pour la MA et 0,08 μT pour la MG. En supprimant les CM enregistrés pendant le sommeil, ces moyennes deviennent respectivement 1,04 et 0,08 μT.

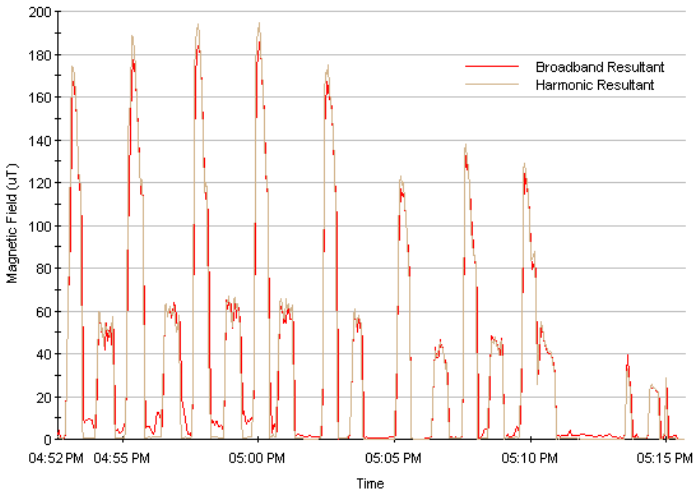

FIG. 4.11 – CM enregistrés dans un train par un passager.

4.3 Comparaison des expositions moyennes

Pour comparer les moyennes observées par deux groupe de populations, nous utilisons des tests statistiques qui cherchent à déterminer si K ($K \geq 2$) échantillons peuvent être considérés comme étant issus d'une même population. Ces tests sont aussi appelés *tests d'homogénéité*.
Pour réaliser les tests, nous allons nous servir des moyennes arithmétiques et géométriques. Du fait que leurs lois ne sont pas connues, on ne peut pas se contenter des tests paramétriques classiques. Une alternative est d'appliquer des tests non paramétriques basés sur les rangs des observations.

On parle de test non paramétrique lorsque l'on ne fait aucune hypothèse sur la distribution des variables. On parle aussi de *distribution free*, c'est-à-dire que la qualité des résultats ne dépend pas, a priori, de la distribution sous-jacente des données. Il existe différents types de tests non paramétriques,

le choix dépend des questions auxquelles on veut répondre. Le test le plus utilisé est le test de Kolmogorov-Smirnov. C'est un test de rang qui permet de décider, à partir d'une variable d'intérêt continue, si deux échantillons peuvent être considérés comme étant issus de la même population ou non. Cela revient à tester l'égalité des fonctions de répartition empirique de la variable d'intérêt. Toute différence conduit au rejet de cette hypothèse. Les différences sont de deux types : elles peuvent être liées à des dispersions (paramètres d'échelle) ou des paramètres de tendance centrale (caractéristique de localisation) non identiques statistiquement dans les deux échantillons.
Le second cas est celui qui nous intéresse. On aimerait savoir si il y a un décalage entre les fonctions de répartition de plusieurs sous-populations dû à un caractère de localisation.

4.3.1 Test de rang dans un modèle de localisation

Les tests de rang sont principalement destinés à détecter les écarts entre deux fonctions de répartition sans en préciser la nature. Les sources principales de différenciation sont de deux types : les modèles de localisation qui stipulent que l'écart est attribué au décalage entre les caractéristiques de tendance centrale des distributions ; les modèles d'échelle indiquant que l'écart est dû à des dispersions différentes.
Ces tests peuvent être rapprochés avec les tests paramétriques de comparaison de moyennes ou de variances dans le cas de données gaussiennes. L'avantage des tests non paramétriques est que cette contrainte est levée et le champ d'application est plus large. Il suffit que la distribution des données soit continue pour que les tests soient applicables.

La transformation des données en rangs introduit une propriété très appréciable lorsqu'on traite des problèmes réels : les statistiques sont moins sensibles aux observations aberrantes. En fait la présence de points atypiques fausse très souvent la moyenne qui joue un rôle central dans les tests paramétriques. Si le point atypique correspond à une très grande valeur, s'écartant fortement des autres, la moyenne est tirée vers le haut, biaisant ainsi les calculs subséquemment. Avec les rangs, on utilise l'information « le point correspond à la valeur la plus élevée » : le rôle néfaste du point atypique est amoindri.

Dans le cas ici présent, nous nous intéressons à des modèles de localisation. Toutefois, nous vérifierons, par le test robuste de Moses, que les paramètres d'échelle sont égaux sur les sous-populations. On entend par paramètre de localisation, un estimateur robuste de tendance centrale. Le mot robuste s'explique par la tolérance par rapport aux valeurs extrêmes ou atypiques. Pour cela, nous utilisons le test de Wilcoxon-Mann-Whitney dans le cas de deux échantillons et celui de Kruskal-Wallis pour plus de

deux échantillons comme lorsque nous voulons tester l'homogénéité des CM moyens enregistrés dans les foyers proches des réseaux à haute tension, dans les foyers à côtés des réseaux ferrés électrifiés et ceux éloignés de ces ouvrages. Si les paramètres d'échelle ne sont pas égaux, ces deux tests ne sont pas applicables. Dans ces conditions, nous appliquons le test robuste de Fligner-Policelo.

4.3.1.1 Test robuste de Moses

Le test robuste de Moses est un test non paramétrique qui se distingue principalement des autres par le fait qu'il ne requiert pas l'égalité des paramètres de localisation pour tester l'égalité des paramètres d'échelle. Il ne repose pas tout à fait sur une statistique de rang linéaire. Il utilise le test de Wilcoxon-Mann-Whitney pour établir le rejet ou non de l'hypothèse nulle. Sa mise en œuvre nécessite des calculs sur les données brutes pour produire des valeurs intermédiaires que l'on présente au test de Wilcoxon-Mann-Whitney. Il ne fonctionne que sur des données quantitatives. En anglais, il est connu sous l'appellation « Moses rang-like test for scale differences ».

La réalisation du test se fait en deux étapes :

1. **Construction des données intermédiaires**
   A partir des données initiales, on construit des données intermédiaires qui traduisent la variabilité dans les sous-populations. Les données sont subdivisées en sous groupes de taille h ($h \geq 2$) de telle sorte que les individus inclus dans un groupe appartiennent à la même sous-population. Pour chaque bloc, nous introduisons un indicateur de sa variabilité interne en calculant la somme des écarts à la moyenne du bloc. Pour le bloc l, la formule est donnée par (4.1).

   $$\tilde{x_l} = \sum_{j=1}^{h} (x_{jl} - \bar{x_l})^2 \tag{4.1}$$

   où $\bar{x_l}$ est la moyenne de la variable d'intérêt dans le bloc l.

   Nous disposons ainsi de m nouvelles observations synthétiques dont m_1 sont relatives à la première sous-population et m_2 à la seconde, dans le cas de deux sous échantillons. Si la dispersion est plus forte (respectivement moins forte) dans la première sous-population, on s'attend à ce qu'en moyenne, les valeurs $\tilde{x_l}$ associées soient plus élevées (respectivement plus faibles).
2. **Application du test de Wilcoxon-Mann-Whitney**
   Partant de l'idée que les valeurs intermédiaires $\tilde{x_l}$ sont des indicateurs de variabilité, et qu'elles sont associées chacune à une sous-population,

cette étape consiste à comparer les caractéristiques de tendance centrale de la distribution des $\tilde{x_l}$ dans les sous populations. On peut utiliser un test non paramétrique comme celui de Wilcoxon-Mann-Whitney. Dans ce cas, le schéma du test de Wilcoxon-Mann-Whitney n'est pas modifié sauf que nous disposons de $m = m_1 + m_2$ observations. Ce sont les valeurs $\tilde{x_l}$ qui sont transformées en rangs. La définition de la statistique de test et de la région de rejet est strictement la même.

4.3.1.2 Test Wilcoxon-Mann-Whitney

Si on note $F_1(X)$ et $F_2(X)$, les fonctions de répartition de la variable d'intérêt X respectivement dans les sous-populations 1 et 2, le test se formalise par :

$$H_0 : F_1(X) = F_2(X + \theta), \theta = 0$$
$$H_1 : F_1(X) = F_2(X + \theta), \theta \neq 0$$

Il est aussi possible de spécifier des tests unilatéraux. Pour traduire l'idée « X prend stochastiquement des valeurs plus faibles (respectivement plus élevées) dans le premier échantillon », nous précisons $H_1 : \theta < 0$, (respectivement, $H_1 : \theta > 0$). La manière dont nous avons écrit le test d'hypothèses stipule que l'écart en X est constant tout au long des fonctions de répartition.

La figure 4.12 est un exemple de comparaison de deux fonctions de répartition empiriques de données simulées.

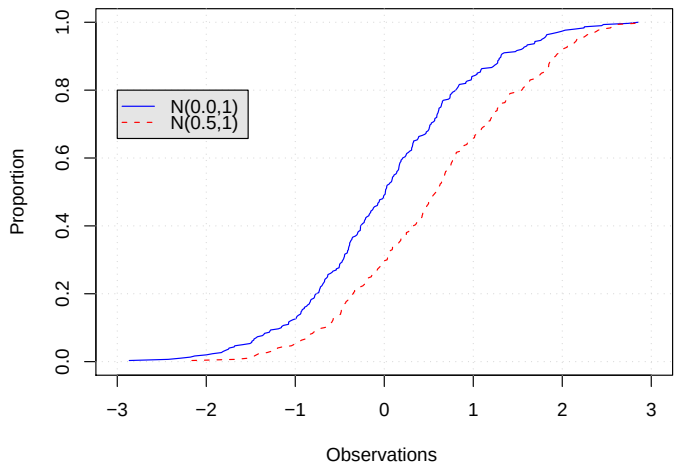

FIG. 4.12 – Décalage entre deux fonctions de répartition de deux lois normales de moyennes $\mu_1 = 0$ et $\mu_2 = 0,5$ et de variances $\sigma_1^2 = \sigma_2^2 = 1$.

Le test de Wilcoxon-Mann-Whitney est le plus connu des tests non paramétriques de localisation. Il existe deux formulations qui se déduisent l'une de l'autre, le test de Wilcoxon d'une part, le test de Mann-Whitney d'autre part. Ce test est basé sur la somme des rangs des observations. Nous définissons ci-dessus les outils nécessaires pour la mise en œuvre du test.

1. **Rangs et somme des rangs :**
   Du fait que le test soit basé sur les rangs, on transforme les observations en rangs c'est-à-dire qu'à la valeur x_i de l'individu numéro i correspond son rang r_i dans les deux sous-échantillons réunis. Pour spécifier son groupe d'appartenance k ($k = 1, 2$), à la valeur x_{ik} correspond son rang r_{ik}. Cette formulation est toutefois trompeuse car le rang est associé à l'ensemble des données (les deux sous échantillons réunis). On peut alors former la somme des rangs pour chaque sous échantillon par :

$$R_k = \sum_{i=1}^{n_k} r_{ik} \tag{4.2}$$

   où n_k est la taille de l'échantillon k.
   Cette transformation induit deux conséquences importantes :
   – La distribution des nouvelles données (les rangs) devient nécessairement symétrique quelque soit la distribution initiale des observations.
   – Le rôle des observations atypiques est considérablement amoindri.
   Même si les observations sont continues, il n'est pas exclu qu'il y ait des ex-aequo. Cela demande un traitement particulier.

2. **Traitement des ex-aequo et le principe de rang moyen**
   Lorsqu'il y a des ex-aequo dans les observations, nous utiliserons la méthode de rang moyen. Les observations possédant des valeurs identiques se voient attribuer la moyenne de leurs rangs. Cette approche est plus puissante mais la variance des statistiques de test est modifiée. On précisera la nature de la correction à introduire lors du calcul des lois et des statistiques de test.
   Il existe une autre méthode qui consiste à attribuer aléatoirement les rangs aux observations confondues. Dans ce cas, aucune modification des tables et lois asymptotiques existantes n'est nécessaire. Cependant, d'une part, la puissance du test est plus faible que celle de la méthode précédente ; d'autre part, la possibilité que la conclusion puisse être différente d'un coup à l'autre ne peut pas être exclue.
   Une fois ce traitement réalisé, on définit le rang moyen du groupe k par :

$$\bar{R}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} r_{ik}. \tag{4.3}$$

3. **Statistiques de rang linéaires**
   Le test de Wilcoxon-Mann-Whitney s'appuie sur une statistique de rang linéaire. Pour cela, on doit choisir un groupe de référence, par convention le premier. C'est celui qui a la plus petite taille n_1 ($n_1 < n_2$, où n_2 est la taille du second groupe). Dans le cas d'une égalité, le groupe de référence est celui qui a la plus petite somme des rangs ($R_1 < R_2$). Une statistique de rang linéaire s'écrit [30, 31] :

$$T = \sum_{i=1}^{n_1} f(r_{i1}) \tag{4.4}$$

   où $f(r_i)$ s'appelle code ou score et $f(.)$ est une fonction score.
   Dans la littérature, on retrouve aussi une autre écriture (4.5) équivalente à (4.3).

$$T = \sum_{i=1}^{n} c_i \times f(r_i) \tag{4.5}$$

   où c_i est une fonction indicatrice qui vaut 1 si l'individu i appartient au groupe de référence et 0 sinon.

   La statistique T possède une propriété très importante : elle converge vers une loi normale lorsque les tailles des deux échantillons augmentent [32, 33]. Le choix de la fonction $f(.)$ dépend des informations que l'on veut mettre en avant.

Pour le test de Wilcoxon-Mann-Whitney, la fonction $f(.)$ correspond à la fonction identité et les scores correspondent directement aux rangs. Dans ce cas, la statistique de test est notée W_s et est égale à la somme des rangs du groupe de référence (4.6).

$$W_s = \sum_{i=1}^{n_1} r_{i1} \tag{4.6}$$

Théorème 4.1 *Sous l'hypothèse H_0, on peut calculer l'espérance et la variance de la variable W_s. Elles sont données par (4.7) et (4.8).*

$$\mathbb{E}(W_s) = \frac{1}{2} n_1 (n_1 + n_2 + 1). \tag{4.7}$$

$$\mathbb{V}(W_s) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1). \tag{4.8}$$

La demonstration de (4.7) et (4.8) est donnée dans l'annexe "Compléments du chapitre 4".

Mann-Whitney [32] et Hoeffding [33] ont montré que la variable aléatoire Z définie par $Z = \frac{W_s - \mathbb{E}(W_s)}{\sqrt{\mathbb{V}(W_s)}}$ converge vers une loi normale centrée et réduite lorsque n_1 et n_2 sont assez grands. Autrement dit :

$$\lim_{\min(n_1,n_2)\to+\infty} P\left\{\frac{W_s - \mathbb{E}(W_s)}{\sqrt{\mathbb{V}(W_s)}} < t\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{1}{2}x^2}dx. \qquad (4.9)$$

Pour un test bilatéral ($H_0 : F_1(X) = F_2(X)$ contre $H_1 : F_1(X) = F_2(X+\theta)$, avec $\theta \neq 0$), on rejette l'hypothèse nulle si la valeur absolue de Z observée est supérieure au quantile $1-\frac{\alpha}{2}$ d'une loi normale centrée et réduite avec α le seuil de significativité du test. En d'autres termes, l'hypothèse H_0 est rejetée si la p-value est inférieure à α. La p-value est ici la probabilité que la valeur absolue d'une variable aléatoire de loi normale centrée et réduite soit supérieure à la valeur absolue de Z observée.

Pour un test unilatéral à gauche (respectivement à droite), on rejette l'hypothèse nulle si la valeur de Z observée est inférieure (respectivement supérieure) au quantile $1-\alpha$ d'une loi normale centrée et réduite.

Remarques :

1. **Correction pour les ex-aequo**
   Les tests non paramétriques ne reposent pas sur des hypothèses concernant la distribution sous-jacente des données. En revanche, on suppose que la fonction de répartition est continue. Lorsque des ex-aequo sont associés à des individus de même groupe, la statistique de test n'est pas modifiée. Par contre, lorsque des individus de groupes différents présentent la même valeur et se voient attribuer des rangs (moyens) identiques, la statistique du test est modifiée par rapport à la méthode des rangs moyens. Ainsi lorsqu'on veut utiliser l'approximation normale pour définir la région critique du test, la variance de la statistique doit être corrigée. Elle devient [34] :

$$V_{W_s} = \mathbb{V}(W_s)\left(1 - \frac{1}{n^3 - n}\sum_{j=1}^{n_0}\widetilde{n}_j(\widetilde{n}_j^2 - 1)\right) \qquad (4.10)$$

   où $n = n_1+n_2$ est l'effectif total, n_0 est l'ensemble des valeurs distinctes dans les deux sous échantillons réunis, $\widetilde{n}_j$ est le nombre d'observations associé à la valeur numéro j.

2. **Correction de continuité**
   Lorsque les effectifs sont de taille modérée, on peut améliorer l'approximation normale en introduisant la correction de continuité [35]. Cette modification est nécessaire lorsqu'on veut faire une approximation de la loi d'une variable discrète à une loi continue comme dans le cas d'une loi binomiale ou de Poisson.
   Pour un test bilatéral, la statistique de test est donnée par :

$$|Z| = \frac{|W_s - \mathbb{E}(W_s)| - 0.5}{\sqrt{\mathbb{V}(W_s)}}. \tag{4.11}$$

   Et la règle de décision n'est pas modifiée.

   Pour les tests unilatéraux, on introduit simplement la correction de continuité. La statistique est donnée en 4.12 :

$$Z = \frac{W_s - \mathbb{E}(W_s) \pm 0.5}{\sqrt{\mathbb{V}(W_s)}}. \tag{4.12}$$

   Pour un test unilatéral à gauche, nous rajoutons 0,5, à droite nous retranchons 0,5. Les règles de décisions restent toutefois inchangées.

Le test de Wilcoxon-Mann-Whitney a en général un bon comportement. Si la distribution des données est gaussienne, il est un peu moins puissant que le test (paramétrique) de Student. Dans les autres cas, il le surclasse [30].

4.3.1.3 Test de Kruskal-Wallis

Les tests non paramétriques de comparaison de population peuvent être étendus à K populations ($K > 2$), tout comme le test de Student de comparaison de moyennes peut être généralisé en analyse de la variance permettant de comparer simultanément les moyennes de K échantillons gaussiens. On définit en quelque sorte une analyse de la variance sur les rangs, ou plus précisément sur les scores déduits des rangs.

La formulation du test consiste à savoir si les fonctions de répartition conditionnelles $F_k(X)$ sont toutes identiques en supposant que les paramètres d'échelle sont égaux deux à deux. L'hypothèse nulle s'écrit :

$$H_0 : F_1(X) = F_2(X) = \cdots = F_k(X) = \cdots = F_K(X)$$

L'hypothèse alternative correspond à « au moins une fonction $F_k(X)$ diffère des autres ».

Le test de Kruskal-Wallis est la généralisation à K populations du test bilatéral de la somme des rangs de Wilcoxon-Mann-Whitney. Il est considéré

comme l'alternative non paramétrique de l'ANOVA, dès que la distribution des données n'est pas gaussienne.
Le rapprochement avec l'analyse de la variance est justifiée jusque dans la construction de la statistique de test. Soit $\bar{R}$ la moyenne globale des rangs, et $\bar{R}_k$ la moyenne des rangs pour les observations du groupe k, la statistique de Kruskal-Wallis est définie en (4.13) [35, 31, 36] :

$$H = \frac{12}{n\,(n+1)} \sum_{k=1}^{K} n_k \left(\bar{R}_k - \bar{R}\right)^2 \tag{4.13}$$

C'est l'expression d'une variabilité inter-classes, c'est-à-dire la dispersion des moyennes conditionnelles autour de la moyenne globale. Si l'hypothèse nulle est vérifiée, les moyennes conditionnelles des rangs sont proches de la moyenne globale : H prend une valeur proche de 0. La région critique correspond aux grandes valeurs de H. Plus H s'écarte de 0, plus l'hypothèse alternative sera crédible.

Sous H_0, H suit une loi de χ^2 à $K-1$ degrés de liberté. L'hypothèse nulle est rejetée si la valeur observé de H est supérieur au quantile $1-\alpha$ d'une loi de $\chi^2(K-1)$.

En effet, $\forall k \in \{1, 2, \cdots K\}$, on pose : $H_k = \frac{12}{n(n+1)} n_k \left(\bar{R}_k - \bar{R}\right)^2$

$H_k = \frac{n_k\left(\bar{R}_k - \frac{n+1}{2}\right)^2}{\frac{n(n+1)}{12}}$ car $\bar{R} = \frac{n+1}{2}$.

En remplaçant$\bar{R}_k$ par $\frac{R_k}{n_k}$, on trouve $H_k = \frac{\left(R_k - \frac{n_k(n+1)}{2}\right)^2}{\frac{n_k n(n+1)}{12}}$.

Sous H_0 et pour n_k assez grand, R_k est une variable aléatoire qui converge approximativement vers une loi normale d'espérance et de variance respectives :

$$\mathbb{E}\left(R_k\right) = \frac{n_k(n+1)}{2}$$

$$\mathbb{V}\left(R_k\right) = \frac{n_k n\,(n+1)}{12}$$

Ainsi, sous ces hypothèses, H_k est une variable aléatoire qui suit une loi de $\chi^2(1)$. Et pour finir, $H = \sum_{k=1}^{K} H_k$ suit une loi de χ^2 à $K-1$ degrés de liberté car les R_k sont reliées par une relation linéaire.

Lorsque les données comportent des ex-aequo, nous utilisons le principe du rang moyen et la statistique du test devra être corrigée. En notant n_0 le nombres d'observations distinctes dans l'ensemble des données, $\tilde{n}_j$ le nombre d'observations pour la valeur numéro j $(j = 1, 2, \cdots, n_0)$, la statistique de test s'écrit :

$$\tilde{H} = \frac{H}{1 - \frac{1}{n^3 - n} \sum_{j=1}^{n_0} \tilde{n}_j (\tilde{n}_j^2 - 1)}. \tag{4.14}$$

Les tests de Wilcoxon-Mann-Whitney et de Kruskal-Wallis décrits dans ces deux derniers paragraphes reposent sur une hypothèse importante qui est l'égalité des dispersions dans les échantillons. Pour pouvoir réaliser ces tests, on doit s'assurer que cette hypothèse est vérifiée. Pour cela, nous utilisons le test robuste de Moses pour vérifier l'égalité des paramètres d'échelle faute de quoi les tests ne peuvent pas être appliqués.

4.3.1.4 Test de Fligner-Policello

Dans le cas ou les tests de Wilcoxon-Mann-Whitney ou de Kruskal-Wallis ne seraient pas réalisables par le fait que le test de Mores rejette l'hypothèse d'égalité des dispersions, nous utilisons le test de rang robuste de Fligner-Policello. Ce test permet de vérifier l'égalité des caractères de localisation de deux sous-populations lorsque la variable d'intérêt n'est pas gaussienne avec des dispersions différentes sur les deux échantillons.
Pour le définir, on note :

$$P_i = \sum_{j=1}^{n_2} \mathbb{I}_{x_{i1} > x_{j2}} \tag{4.15}$$

$$\bar{P} = \frac{1}{n_1} \sum_{i=1}^{n_1} P_i \tag{4.16}$$

$$V_P = \sum_{i=1}^{n_1} (P_i - \bar{P})^2 \tag{4.17}$$

avec $n_1 < n_2$ et $\mathbb{I}(x)$ définie en (4.18)

$$\mathbb{I}(x) = \begin{cases} 0 & \text{si } x_{i1} < x_{j2} \\ 1 & \text{si } x_{i1} > x_{j2} \\ 0,5 & \text{si } x_{i1} = x_{j2} \end{cases} \tag{4.18}$$

De la même manière, on note

$$Q_i = \sum_{j=1}^{n_1} \mathbb{I}_{x_{i2} > x_{j1}} \tag{4.19}$$

On calcule aussi $\bar{Q}$ et V_Q comme dans (4.16) et (4.17).
La statistique de Fligner-Policello testant l'égalité des caractéristiques de localisation est définie en [36] est donnée par :

$$U = \frac{\sum_{i=1}^{n_1} P_i - \sum_{i=1}^{n_2} Q_i}{2\sqrt{V_P + V_Q + \bar{P}\bar{Q}}}. \tag{4.20}$$

Pour $(n_1, n_2 > 12)$, U converge vers une loi normale centrée et réduite [36]. Le test rejette l'hypothèse nulle si la valeur absolue de U observée est supérieure au quantile à $1 - \alpha/2$ d'une loi normale centrée et réduite.

4.3.1.5 Mise en garde pour la réalisation des tests

L'application des tests décrits dans ce paragraphe repose sur la convergence des statistiques de rang vers une loi gaussienne. Cette convergence est obtenue lorsque les deux échantillons sont de grandes tailles. C'est ce qu'exigent même les tests, qu'ils soient paramétriques ou non. Lorsque les effectifs sont faibles, les tests paramétriques, ne sont pas utilisables (à moins vraiment que l'hypothèse de normalité soit établie), à la différence des tests non paramétriques. La convergence vers les lois asymptotiques est très rapide pour les tests non paramétriques. Dans la pratique, dès que les effectifs atteignent un niveau modéré (de l'ordre de 20 à 30 observations, selon le test), les approximations sont efficaces. Pour le test de Wilcoxon-Mann-Whitney par exemple, il suffit que $n_1 > 10$ (ou $n_2 > 10$) pour que l'approximation normale soit valable [36]. On propose $n_1 + n_2 > 20$ avec $n_1 > 3$ et $n_2 > 3$ dans [34], ou $n_1 > 8$ et $n_2 > 8$ dans [31].

Le choix de h pour le test robuste de Moses influe sur la qualité des résultats : si h est grand, les indicateurs de variabilité dans les blocs $\tilde{x}_l$ sont de bonne qualité, mais le nombre de valeurs m qu'on présentera au test de Wilcoxon-Mann-Whitney sera faible, mettant en péril la teneur des conclusions de celui-ci. Dans le cas contraire, si on diminue h, m sera grand et le test produira des résultats fiables, mais sur des valeurs $\tilde{x}_l$ estimées sur trop peu d'observations. Nous prendrons $h = E(\sqrt{\min(n_1, n_2)})$ tout en s'assurant que le test Wilcoxon-Mann-Whitney est réalisable en se référant au moins à l'une des références [31, 36, 34] ($E(x)$ désigne la partie entière de x). Le test de Moses utilise un découpage en sous groupes d'échantillons : pour chaque réalisation du test, 100 sous groupes sont tirés aléatoirement. Nous calculons ensuite la p-value moyenne observée ainsi qu'un intervalle de cette dernière de niveau 95%.

La démarche est facilement généralisable à la comparaison des dispersions dans K populations ($K \geq 2$). La constitution des données intermédiaires ne posent aucun problème, et il suffit de substituer le test de Kruskal-Wallis au test de Wilcoxon-Mann-Whitney.

4.3.1.6 Réalisation des tests

Pour comparer les paramètres de localisation de plusieurs échantillons, nous suivons les étapes décrites dans la figure 4.13.

Sans précision de notre part, l'hypothèse nulle est « les paramètres de localisation des échantillons considérés sont identiques ». Pour l'alternative, nous préciserons dans chaque cas de figure pour le test de Wilcoxon-Mann-Whitney alors qu'elle sera « au moins, deux échantillons ont des paramètres de localisation différents », pour celui de Kruskal-Wallis (plus de deux échantillons). En fait, le choix de l'hypothèse alternative dépend de la position des fonctions de répartition empirique l'une par rapport à l'autre. Le seuil de significativité de l'ensemble des tests est fixé à $\alpha = 5\%$.

Les résultats affichés sont :
- La p-value moyenne et l'intervalle de confiance observé avec le test de Moses pour 100 tirages indépendants (4.13).
- La conclusion du test d'égalité des paramètres de dispersion.
- Le test à réaliser pour l'égalité des paramètres de localisation.
- Conclusion finale.

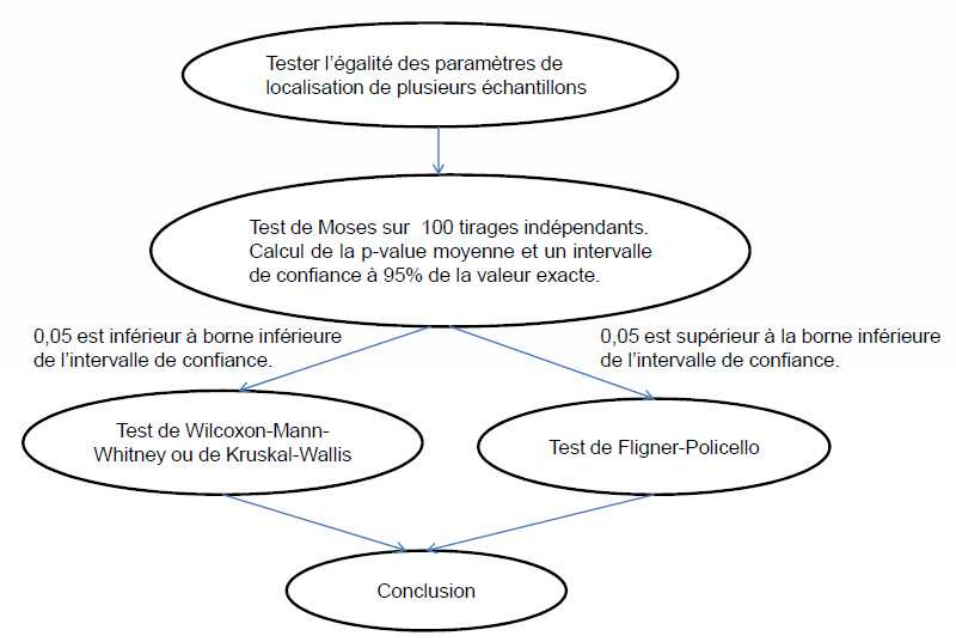

FIG. 4.13 – Organigramme pour la comparaison des paramètres de localisation de plusieurs échantillons.

4.3.2 Comparaison des expositions des enfants et des adultes

Les CM enregistrés dépendent des activités réalisées pendant la période de mesure mais aussi des lieux où elles sont exercées. On peut imaginer que les enfants qui, en période scolaire, restent soit à l'école soit au domicile, sont moins exposés que les adultes qui, eux exercent des activités diverses et variées. Pour vérifier cette hypothèse, des tests de comparaison des moyennes arithmétiques et géométriques des deux populations sont réalisés. Le test utilisé est celui de Wilcoxon-Mann-Whitney et l'échantillon de référence est celui composé des enfants car ils sont moins nombreux que les adultes.

Avec des tailles de 978 individus pour les enfants et 1 054 pour les adultes, le test de Wilcoxon-Mann-Whitney est réalisable moyennant l'égalité des dispersions dans les deux échantillons. Pour vérifier cela, nous appliquons le test robuste de Moses. Pour chaque type de moyenne, des blocs de 31 observations sont réalisés ($h = 31$). Avec ce choix, les échantillons de données intermédiaires auxquelles le test de Wilcoxon-Mann-Whitney est basé pour mesurer la variabilité sont de tailles 31 pour les enfants et 34 pour les adultes.

1. **Exposition sur 24 heures :**
   Les p-values moyennes observées pour le test de Moses sont de 0,076 (IC=[0,060 ; 0,091]) pour les MA et 0,005 (IC=[0,004 ; 0,006]) pour les MG. Comme α n'appartient pas à l'un des intervalles et appartient à l'autre, on peut ainsi considérer que les dispersions sont égales pour les MA et ne le sont pas pour les MG. On applique le test de Wilcoxon-Mann-Whitney pour les MA et celui de Fligner-Policello pour la comparaison de ces moyennes prises deux à deux.
   Les rangs moyens observés pour les MA sont respectivement de 837,8 et 1182,3 pour les enfants et les adultes. Pour les MG, ils sont de 874,8 pour les enfants et 1148,0 pour les adultes. Ces valeurs indiquent que, en moyenne, les enfants sont moins exposés que les adultes. La figure 4.14 donne les fonctions de répartition empirique des MA pour les deux types de populations. Elle montre que celle des enfants est excentrée vers la gauche par rapport à celle des adultes. Autrement dit, pour une proportion donnée, le percentile associé à la MA des enfants est inférieur à celui observé chez les adultes. Pour vérifier cette hypothèse, les tests de Wilcoxon-Mann-Whitney et de Fligner-Policello sont appliqués. Il sont basés sur l'hypothèse stipulant qu'il n'y a pas de différences entre les enfants et les adultes en termes de MA ou de MG contre une exposition moins élevée chez les enfants que chez les adultes. Les p-values observées sont inférieures à 0,001. Les tests rejettent l'hypothèse d'une exposition homogène pour les deux populations en faveur d'une exposition plus élevée pour les adultes.

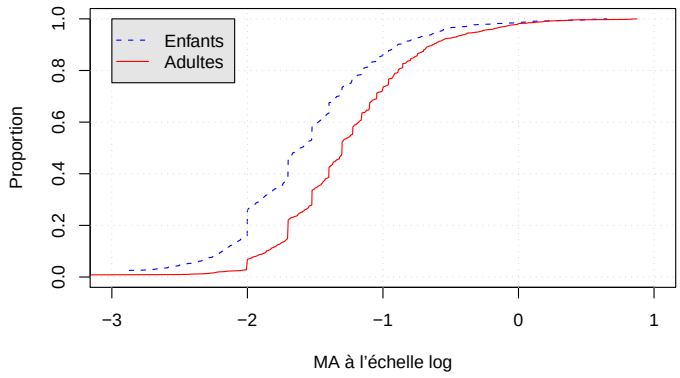

FIG. 4.14 – Fonctions de répartition empirique des MA des enfants des adultes à l'échelle log à base 10.

2. **Exposition hors sommeil :**
   Les tests de Moses réalisés sur les MA et MG des deux populations rejettent l'égalité des dispersions dans chaque type de moyenne. Pour les MA la borne supérieure de l'intervalle de confiance de la p-value est inférieure à 0,001 et pour les MG, la p-value moyenne observée est de 0,011 (IC=[0,009 ; 0,013]). Le test de Wilcoxon-Mann-Whitney rejette aussi l'hypothèse d'homogénéité en faveur d'une exposition plus élevée chez les adultes (les p-values sont inférieures à 0,001).

4.3.3 Comparaison des expositions en Île-de-France et dans les autres régions

La région Île-de-France est la plus peuplée et la plus équipée en matière d'infrastructures électriques (RER, trains, métros, tramway, etc.). On peut ainsi penser que l'exposition est plus élevée dans cette région par rapport aux autres. Sous cette hypothèse, nous pouvons dire que l'exposition de la population française est surestimée car nous avons observé plus de volontaires en Île-de-France par rapport au nombre prévu pour les adultes (tableau 3.4). Pour vérifier cette hypothèse, nous comparons les moyennes arithmétiques et géométriques observées en Île-de-France et dans les autres régions. Les hypothèses nulle et alternative sont respectivement « il n'y a pas de différence entre les expositions moyennes observées en Île-de-France et celles des autres régions » et « les expositions moyennes sont plus élevées en Île-de-France que dans les autres régions ».

4.3.3.1 Les enfants

1. **Exposition sur 24 heures**
   Les p-values moyennes observées pour l'égalité des paramètres de dispersion des MA et des MG dans les deux populations sont respectivement de 0,353 (IC=[0,307 ; 0,400]) et 0,008 (IC=[0,006 ; 0,010]). Ces résultats montrent que pour les MA, les dispersions sont égales alors qu'elles ne le sont pas pour les MG. Pour les tests des paramètres de localisation, nous appliquons le test de Wilcoxon-Mann-Whitney pour les MA et celui de Fligner-Policello pour les MG. Ces tests donnent des p-values inférieures à 0,001 : ils rejettent l'hypothèse nulle en faveur d'une exposition plus élevée en Île-de-France par rapport aux autres régions. Les expositions moyennes observées en Île-de-France sont respectivement de 0,13 μT (IC=[0,06 ; 0,20]) pour les MA et 0,04 μT (IC=[0,03 ; 0,05]) pour les MG. Pour les autres régions, elles sont de 0,07 μT (IC=[0,06 ; 0,09]) pour les MA et 0,01 μT μT (IC=[0,012 ; 0,017] pour les MG.

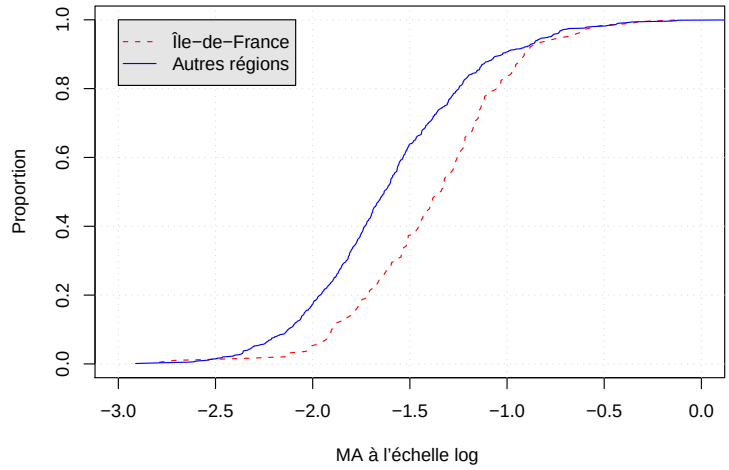

FIG. 4.15 – Fonctions de répartition empirique des MA observées par les enfants hors période de sommeil en Île-de-France et dans les autres régions à l'échelle log à base 10.

2. **Exposition hors période de sommeil**
   Comme pour le cas précédent, les dispersions sont égales pour les MA et différentes pour les MG. Les p-values moyennes pour le test d'égalité de ces dernières sont de 0,369 (IC=[0,324 ; 0,414]) pour les MA et 0,022 (IC=[0,017 ; 0,027]) pour les MG. Nous appliquons respectivement les

tests de Wilcoxon-Mann-Whitney et de Fligner-Policello pour comparer les paramètres de localisation des MA et des MG. Ces tests rejettent l'hypothèse nulle avec des p-values inférieures à 0,001 en faveur d'une exposition plus élevée en Île-de-France par rapport aux autres régions. La figure 4.15 montre le décalage entre les fonctions de répartition empirique des MA observée hors période de sommeil par les enfants. Les moyennes observées en Île-de-France sont de 0,07 μT (IC=[0,06 ; 0,08]) pour les MA et 0,03 μT (IC=[0,02 ; 0,04]) pour les MG. Pour les autres régions, elles sont de 0,05 μT (IC=[0,04 ; 0,06]) pour les MA et 0,02 μT (IC=[0,01 ; 0,03]).

4.3.3.2 Les adultes

1. **Exposition sur 24 heures**
   Le test de Moses d'égalité des dispersions des moyennes observées en Île-de-France et dans les autres régions donne des p-values moyennes de 0,670 (IC=[0,629 ; 0,710]) pour les moyennes arithmétiques et 0,211 (IC=[0,185 ; 0,236]) pour les moyennes géométriques. Ces résultats laissent supposer que les dispersions sont égales pour chaque type de moyenne dans les deux population à 95%. Pour la comparaison des paramètres de localisation, nous appliquons le test de Wilcoxon-Mann-Whitney. Ce test donne des p-values inférieures à 0,001 pour les deux moyennes. Autrement dit, il rejette l'hypothèse nulle en faveur d'une exposition plus élevée en Île-de-France pour les deux moyennes. La figure 4.16 illustre ces résultats pour les moyennes arithmétiques. La fonction de repartition des moyennes arithmétiques observées dans la région parisienne est excentrée vers la droite de celle des MA enregistrées dans les autres régions. Les moyennes observées en Île-de-France sont de 0,17 μT (IC=[0,09 ; 0,25]) pour les MA et 0,05 μT (IC=[0,05 ; 0,06]). Pour les autres régions, les moyennes sont de 0,14 μT (IC=[0,11 ; 0,16]) pour les MA et 0,03 μT (IC=[0,02 ; 0,034]) pour les MG.

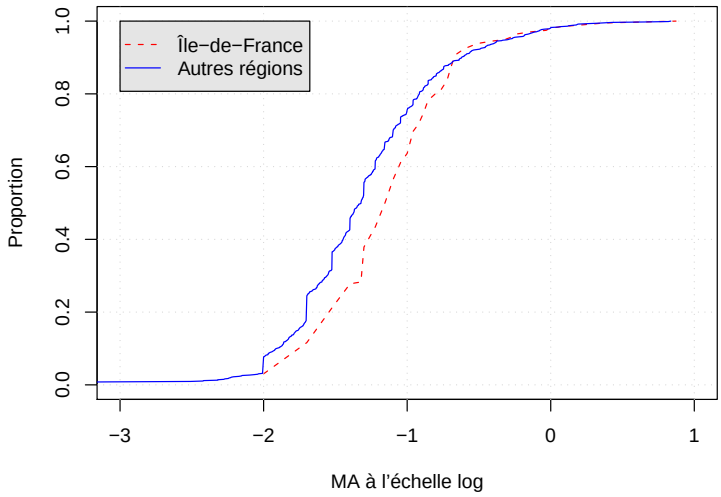

FIG. 4.16 – Fonctions de répartition empirique des MA observées par les adultes sur 24 heures en Île-de-France et dans les autres régions à l'échelle log à base 10.

2. **Exposition hors période de sommeil**
   Les tests d'égalité des dispersions ne rejettent pas cette hypothèse pour les deux types de moyennes. Les p-values moyennes sont de 0,290 (IC=[0,255 ; 0,324]) pour les MA et 0,159 (IC=[0,139 ; 0,178]) pour les MG. Pour réaliser le test de comparaison des paramètres de localisation, nous appliquons le test de Wilcoxon-Mann-Whitney. Ce test rejette l'hypothèse nulle en faveur d'une exposition plus élevée en Île-de-France par rapport aux autres régions avec des p-values inférieures à 0,001 pour les deux types de moyennes. Les moyennes enregistrées en Île-de-France sont de 0,12 μT (IC=[0,09 ; 0,14]) pour les MA et 0,04 μT (IC=[0,03 ; 0,05]) pour les MG. Pour les autres régions, elles sont de 0,10 μT (IC=[0,08 ; 0,12]) pour les MA et 0,03 μT (IC=[0,02 ; 0,04]) pour les MG.

4.3.4 Comparaison des moyennes observées au domicile et à l'extérieur

4.3.4.1 Les enfants

En considérant les expositions sur 24 heures ou hors période de sommeil, le test de Moses rejette l'hypothèse d'égalité des dispersions entre les moyennes au domicile et à l'extérieur et ceci pour les deux types de moyennes (MA et MG). Les p-values moyennes sont inférieures à 0,001

pour les expositions sur 24 heures (IC=$[2,9 \times 10^{-5}\,;\ 1,2 \times 10^{-4}]$ pour les MA et IC=$[6,2 \times 10^{-5}\,;\ 1,2 \times 10^{-4}]$ pour les MG). Elles sont de 0,029 (IC=[0,023 ; 0,035]) et 0,036 (IC=[0,027 ; 0,045]) respectivement pour les moyennes arithmétiques et géométriques hors période de sommeil. Dans tous les cas c'est le test de Fligner-Policello qui est appliqué pour tester l'existence d'un éventuel décalage entre les fonctions de répartition. Nous avons considéré une exposition plus élevée au domicile par rapport à l'extérieur pour l'hypothèse alternative. Les conclusions de ces tests indiquent que, quelque soit les CM considérés, les enfants sont plus exposés au domicile qu'à l'extérieur avec des p-values inférieures à 0,001. Ces résultats peuvent être expliqués par le fait que les enfants sont généralement chez eux ou à l'école. C'est plutôt au domicile qu'ils peuvent être en contact avec des sources de CM (travail sur ordinateur, jeux vidéo, ...). La figure 4.17 montre le décalage entre les fonctions de répartition empirique des MA calculés sur les CM enregistrés à l'extérieur et au domicile hors période de sommeil. Le décalage est faible mais suffisant pour être détecté par le test. C'est une des qualités de ce test.

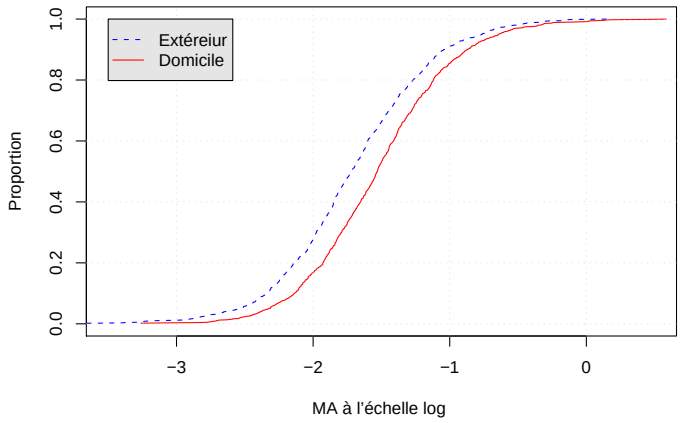

FIG. 4.17 – Fonctions de répartition empirique des MA observées par les enfants à l'extérieur et au domicile hors période de sommeil à l'échelle log à base 10.

4.3.4.2 Les adultes

Pour les adultes, les résultats du test de Moses dépendent des CM considérés. Pour les CM sur 24 heures, le test rejette l'égalité des dispersions pour les deux types de moyennes (les p-values moyennes sont de 0,006, IC=[0,004 ; 0,010] pour les MA et 0,002, IC=[0,001 ; 0,003] pour les MG). Pour les CM hors période de sommeil, le test ne rejette pas cette hypothèse

(les p-values moyennes sont respectivement de 0,229, IC=[0,198 ; 0,260] et 0,607, IC=[0,565 ; 0,648] pour les MA et les MG). Ce qui implique que nous appliquons le test de Fligner-Policello dans le premier cas et celui de Wilcoxon-Mann-Whitney dans le second. Pour cette population, nous prenons comme hypothèse alternative « l'exposition est plus élevée à l'extérieur qu'au domicile ». Ces tests rejettent l'hypothèse d'une exposition identique en faveur d'une exposition plus importante à l'extérieur qu'au domicile (les p-values de ces tests sont inférieures à 0,001). Ces résultats sont totalement le contraire de ceux obtenus chez les enfants. En fait, les principales sources d'exposition des adultes sont plutôt à l'extérieur du domicile (transports ferroviaires, activités professionnelles, ...). Les figures 4.18 et 4.19 donnent les fonctions de répartition empirique des MA observées par les adultes à l'extérieur et au domicile en considérant les CM hors période de sommeil et sur 24 heures. Elles montrent qu'en considérant les CM hors sommeil, l'écart entre les fonctions de répartition est pratiquement constant mais faible. Pour les CM sur 24 heures, on observe un croisement des fonctions de répartition au voisinage du quantile à 90%. Ce résultat peut être dû aux CM générés par les radio-réveils.

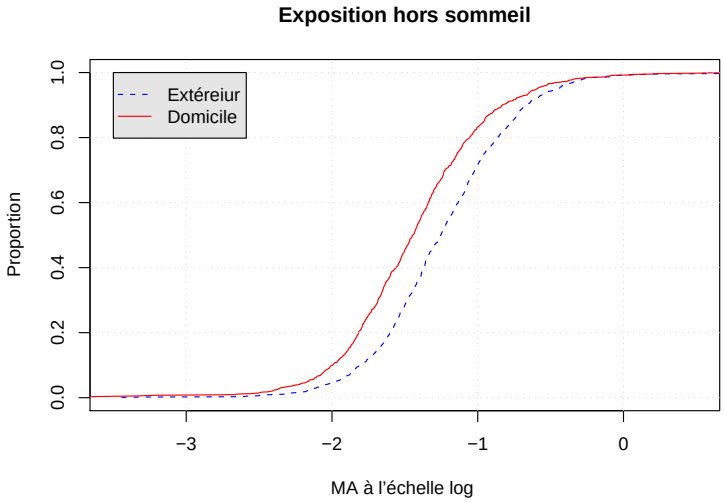

FIG. 4.18 – Fonctions de répartition empirique des MA observées par les adultes à l'extérieur et au domicile hors période de sommeil à l'échelle log à base 10.

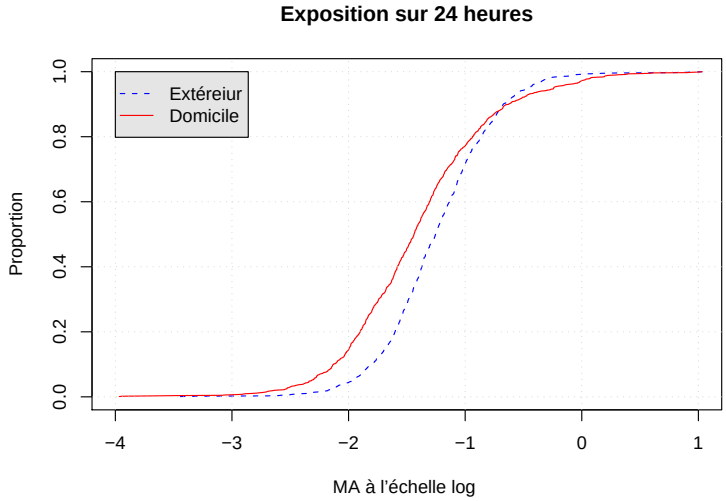

FIG. 4.19 – Fonctions de répartition empirique des MA observées par les adultes à l'extérieur et au domicile avec période de sommeil à l'échelle log à base 10.

4.3.5 Comparaison des moyennes observées au domicile le jour et la nuit

Au foyer, nous exerçons des activités diverses et variées et on peut imaginer que l'exposition en période d'activité n'est pas homogène à celle observée pendant le sommeil. Pour vérifier cela, des tests sont réalisés. Ils sont basés sur les facteurs de localisation des fonctions de répartition. La question est de savoir s'il y a un écart entre les fonctions de répartition des moyennes enregistrées au domicile pendant les périodes d'activité et de sommeil. Pour cela, nous utilisons le test de Wilcoxon-Mann-Whitney ou celui de Fligner-Policello selon la conclusion du test d'égalité des dispersions. Pour réaliser ce dernier, nous choisissons $h = 31$.

4.3.5.1 Les enfants

Les résultats du test de Moses ont permis de conclure que les dispersions ne sont pas égales pour les MA (la p-value moyenne est de 0,005 (IC=[0,003 ; 0,008]) pour les MA et 3×10^{-6} (IC=$[10^{-7}$; $7 \times 10^{-6}]$)) pour les MG. En privilégiant une exposition plus élevée le jour que la nuit au domicile, le test de Fligner-Policello donne des p-values très signifivatives (la p-value est inférieure à 0,001 pour les MA et égale à 0,004 pour les MG). On conclut globalement qu'au domicile, ces moyennes sont plus importantes le jour que la nuit. La figure 4.13 donne les fonctions de répartition empirique des MA enregistrées le jour et la nuit. Elle montre plus particulièrement que l'écart

entre ces fonctions n'est pas constant. Ce problème peut être expliqué par les CM générés par les radio-réveils. On voit bien dans cette figure que les quantiles les plus extrêmes sont plus élevés la nuit que le jour.

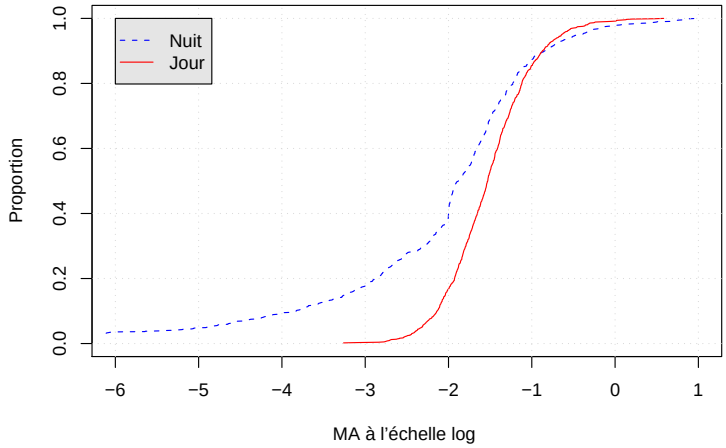

FIG. 4.20 – Fonctions de répartition empirique des MA observées par les enfants au domicile le jour et la nuit à l'échelle log à base 10.

4.3.5.2 Les adultes

En appliquant le même principe, les dispersions ne peuvent pas être considérées comme étant égales pour les deux types de moyennes (les p-values moyennes sont de 5×10^{-6}, IC=$[7 \times 10^{-7}\,;\, 9 \times 10^{-6}]$ pour les MA et 10^{-14}, IC=$[10^{-13}\,;\, 7 \times 10^{-13}]$ pour les MG. Les hypothèses alternatives dépendent de la moyenne considérée. Pour les MA, nous privilégions une exposition plus élevée le jour que la nuit alors que pour les MG, nous stipulons le contraire comme hypothèses alternatives. Le test Fligner-Policello rejette les hypothèses nulles en faveur d'une exposition plus élevée le jour que la nuit pour les MA et le contraire pour les MG (les p-values sont inférieures à 0,001). La figure 4.21 représente les fonctions de répartition empirique des MG. Elle montre aussi que l'écart n'est pas constant mais il peut être considéré constant à partir d'un certain quantile. Les CM liés aux radio-réveils ne peuvent pas expliquer à eux seuls le changement de situation qu'on observe sur les MG car, seulement 26% des adultes ont déclaré avoir posé l'EMDEX à côté du radio-réveil la nuit. C'est une raison de plus qui fait que nous réalisons l'étude en considérant les CM sur 24 heures et hors période de sommeil.

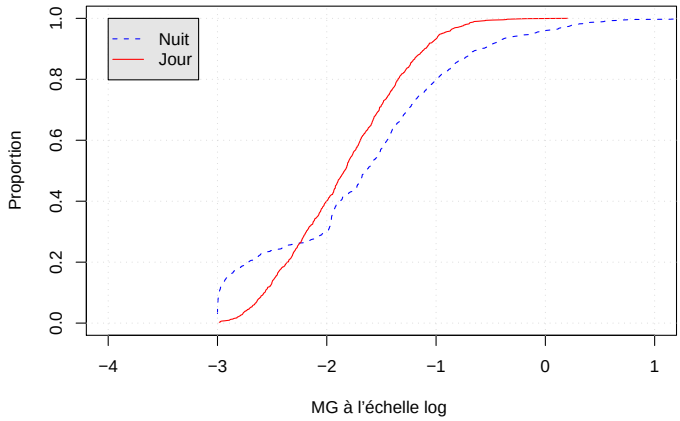

FIG. 4.21 – Fonctions de répartition empirique des MG observées par les adultes au domicile le jour et la nuit.

4.3.6 Comparaison des moyennes par rapport à la proximité des réseaux électriques

Parmi les sources générant des CM, il y a les lignes électriques. On dispose d'informations sur la présence d'ouvrages électriques à proximité des logements. On va s'en servir pour voir s'il y a des différences en termes d'exposition entre les sujets habitant à proximité des lignes de transport d'électricité, des réseaux ferrés électrifiés et les autres. Comme les enfants et les adultes sont différemment exposés, nous réalisons le test sur chaque type de population. Au total, 19 enfants et 23 adultes ont leurs foyers proches de lignes à haute tension. Pour les câbles souterrains, on compte 21 enfants et 16 adultes qui habitent à côté de ces derniers. Différencier les lignes aériennes des câbles souterrains conduirait à un problème de convergence du test de Moses. C'est pour cela que nous ne différencions pas ces deux types d'ouvrages même si nous savons que les profils de champ magnétique généré par ces deux types d'ouvrages sont différents. Nous confondons ces ouvrages et nous les appelons réseaux à haute tension.

Nous commençons par comparer les moyennes observées sur les CM enregistrés au domicile. S'il apparaît des différences, nous réalisons les tests sur les moyennes calculées sur les CM enregistrés sur 24 heures. Du fait qu'on a plus de deux échantillons, le test utilisé est celui de Kruskal-Wallis.

4.3.6.1 Les enfants

Au total, 40 enfants habitent dans des foyers proches des réseaux à haute tension (lignes aériennes ou souterraines). Nous avons aussi identifié 81 enfants qui habitent à côté des réseaux ferrés électrifiés, 5 d'entre eux ont des réseaux à haute tension proches de leurs domiciles. Avec ces tailles d'échantillon, nous avons choisi de faire des blocs de 7 observations pour le test d'égalité des dispersions ($h = 6$). Autrement dit la variabilité est mesurée sur des sous-échantillons de 6 observations et le test est réalisé sur trois sous-échantillons de 6, 13 et 143 données intermédiaires. En se référant à [34], on peut dire que le test de Kruskal-Wallis est applicable.

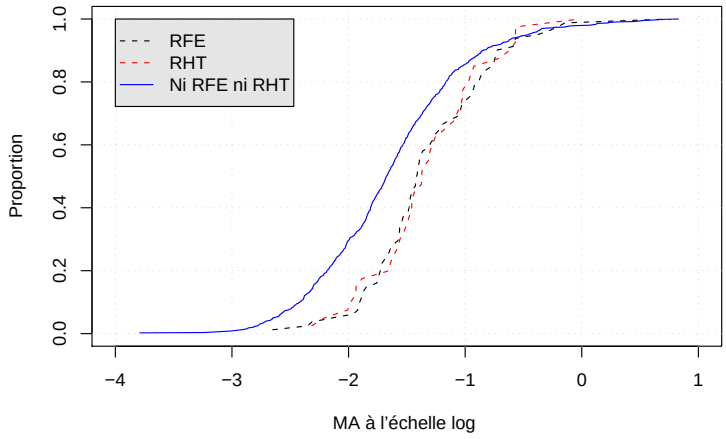

FIG. 4.22 – Fonctions de répartition empirique des MA observées par les enfants dans les foyers proches des réseaux à haute tension (RHT), des réseaux ferrés électrifiés (RFE) et dans les foyers éloignés de ces ouvrages (Ni RFE ni RHT) à l'échelle log à base 10, en considérant les CM au domicile avec période de sommeil.

1. **Moyennes avec période de sommeil :**
   a-Moyennes au domicile avec période de sommeil
   Les tests d'égalité des dispersions dans les trois sous-populations ont donné des p-values de 0,459 (IC=[0,418 ; 0,500]) pour les MA et 0,052 (IC=[0,038 ; 0,066]) pour les MG. On ne rejette pas l'hypothèse d'égalité de ces dispersions pour les MA. Cette hypothèse est par contre rejetée pour les MG. Le test de Kruskal-Wallis appliqué sur les MA pour tester l'égalité des paramètres de localisation dans les trois populations indique que ces paramètres ne sont pas statistiquement identiques (la p-value observée est inférieure à 0,001). En appliquant le test d'égalité

Wilcoxon-Mann-Whitney sur les MA observées dans les foyers proches des réseaux ferrés électrifiés et des réseaux à haute tension, on trouve une p-values de 0,716. Autrement dit, il n'y a pas de différence d'exposition en terme de MA dans les foyers proches des réseaux ferrés électrifiés et des réseaux à haute tension. Dans les foyers éloignés de ces ouvrages, les MA sont moins élevées que dans ces derniers (figure 4.22).

Du fait que les dispersions des MG ne sont pas identiques, nous avons réalisé le test en deux temps :

- Foyers proches des réseaux ferrés et ceux à la fois isolés de ces derniers et des réseaux à haute tension :
  Le test d'égalité des dispersions donne une p-value de 0,093, (IC=[0,068 ; 0,117]). Comme cet intervalle ne contient pas $\alpha = 5\%$, nous concluons que les dispersions sont égales et nous appliquons le test de Wilcoxon-Mann-Whitney pour comparer les paramètres de localisation. Pour l'hypothèse alternative, nous optons pour une exposition plus élevée dans les foyers proches des réseaux ferrés électrifiés. Ce test donne une p-value inférieure 0,001 : les enfant sont plus exposés dans les foyers proches des réseaux ferrés électrifiés que dans les foyers isolés de ces derniers et des réseaux à haute tension.
- Foyers proches des réseaux ferrés électrifiés et ceux situés à proximité des réseaux à haute tension :
  Nous ne rejetons pas l'hypothèse d'égalité des paramètres de dispersion (la p-value moyenne vaut 0,541 et IC=[0,490 ; 0,591]). Pour réaliser le test d'égalité des paramètres de localisation des deux échantillons, nous privilégions une exposition plus élevée dans les foyers situés à proximité des réseaux à haute tension par rapport à ceux situés à côté des réseaux ferrés électrifiés. Avec une p-value de 0,428, le test ne rejette pas l'hypothèse d'égalité des expositions dans les foyers proches des réseaux à haute tension et des réseaux ferrés électrifiés.

Ces résultats montrent que les expositions moyennes sont moins élevées dans les foyers pour lesquels on n'a pas identifié des réseaux à haute tension ou des réseaux ferrés électrifiés à proximité par rapport à ceux situés à côté des ces ouvrages. Pour ces deux types d'ouvrages, les MA ne sont pas statistiquement différentes dans les foyers à proximité alors que les MG sont plus élevées dans les foyers à côté des lignes à haute tension (figure 4.23).

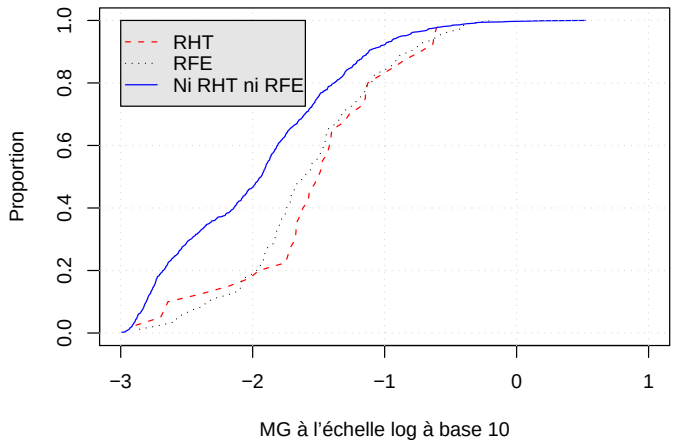

FIG. 4.23 – Fonctions de répartition empirique des MG observées, avec la période de sommeil dans les foyers proches des RHT, des RFE et dans ceux éloignés de ces ouvrages.

b-Moyennes sur 24 heures

Les résultats des tests ont donné les mêmes conclusions que celles tirées sur les moyennes au domicile avec sommeil pour les MA. Il n'apparaît pas de différences entre les MA calculées sur les CM enregistrés chez les enfants habitant près des réseaux à haute tension et des réseaux ferrés électrifiés. Par contre, ces derniers sont plus exposés que ceux habitant dans les foyers éloignés de ces ouvrages. Pour les MG, l'exposition est plus élevée pour les enfants habitant dans les foyers proches des réseaux ferrés électrifiés que dans ceux vivant à côté des réseaux à haute tension. Ces conclusions sont illustrées dans les figures 4.24 et 4.25 donnant les fonctions de répartition empirique des MA et des MG.

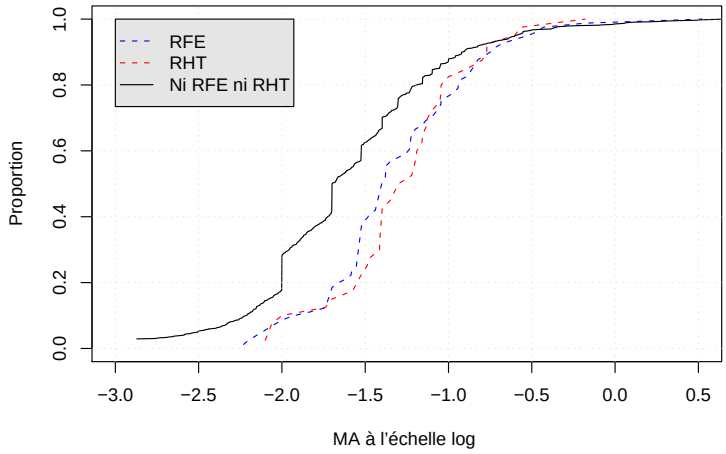

FIG. 4.24 – Fonctions de répartition empirique des MA observées par les enfants habitant proches des réseaux ferrés électrifiés (RFE), des réseaux à haute tension (RHT) et loin de ces ouvrages (ni RFE ni RHT) en considérant les CM sur 24 heures à l'échelle log à base 10.

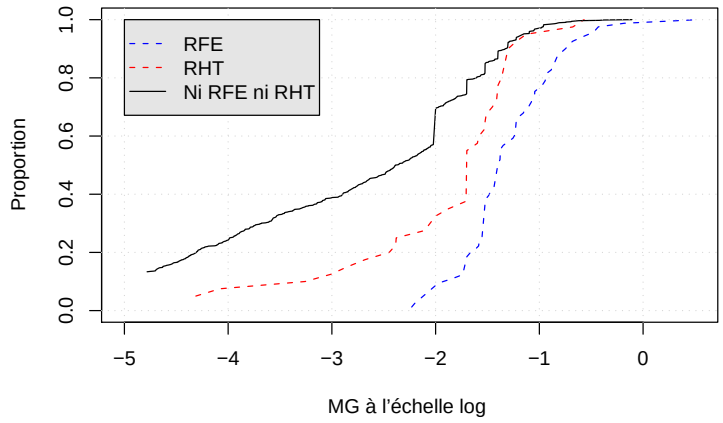

FIG. 4.25 – Fonctions de répartition empirique des MG observées par les enfants habitant proches des réseaux ferrés électrifiés (RFE), des réseaux à haute tension (RHT) et loin de ces ouvrages (ni RFE ni RHT) en considérant les CM sur 24 heures à l'échelle log à base 10.

2. **Moyennes hors sommeil :**

a-Moyennes au domicile hors sommeil

Les conclusions sont les mêmes que dans le cas des moyennes au domicile avec sommeil pour les MA c'est-à-dire qu'il n'y a pas de différence entre les foyers proches des réseaux ferrés électrifiés et les réseaux à haute tension alors que l'exposition est moins élevée dans les foyers isolés des ces ouvrages par rapport à ces derniers. Pour les MG, l'exposition est par contre moins élevée dans les foyers proches des réseaux ferrés électrifiés par rapport au foyers à côté des réseaux à haute tension (figures 4.26 et 4.27).

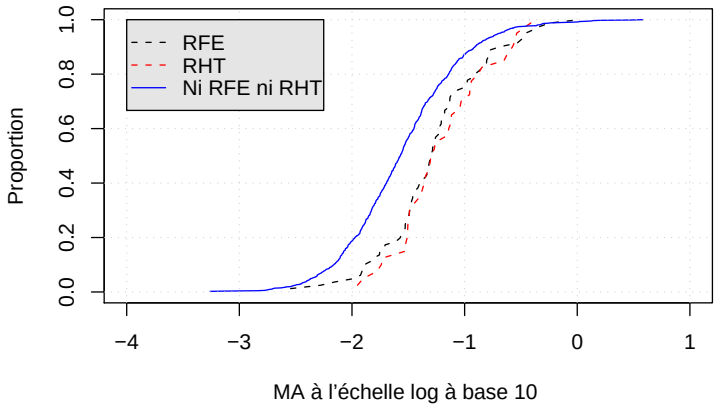

FIG. 4.26 – Fonctions de répartition empirique des MA observées par les enfants habitant proches des réseaux ferrés électrifiés (RFE), des réseaux à haute tension (RHT) et loin de ces ouvrages (ni RFE ni RHT) en considérant les CM au domicile hors période de sommeil.

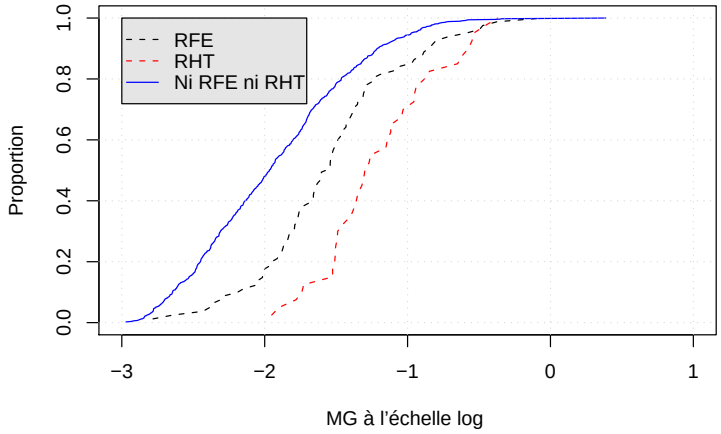

FIG. 4.27 – Fonctions de répartition empirique des MG observées par les enfants habitant proches des réseaux ferrés électrifiés (RFE), des réseaux à haute tension (RHT) et loin de ces ouvrages (ni RFE ni RHT) en considérant les CM au domicile hors période de sommeil.

b-Moyennes sur 24 heures hors sommeil

Les résultats des tests sont les mêmes que en considérant les moyennes sur 24 heures avec période de sommeil (figure 4.28 et 4.29).

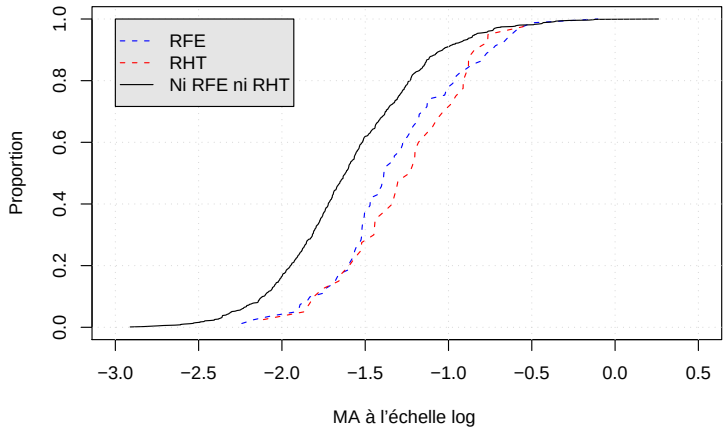

FIG. 4.28 – Fonctions de répartition empirique des MA observées par les enfants habitant proches des réseaux ferrés électrifiés (RFE), des réseaux à haute tension (RHT) et loin de ces ouvrages (Ni RFE ni RHT) en considérant les CM sur 24 heures hors sommeil, à l'échelle log à base 10.

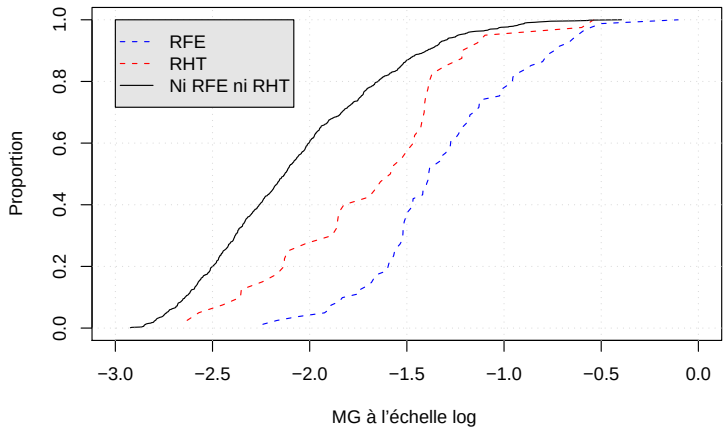

FIG. 4.29 – Fonctions de répartition des MG des enfants habitant proches des réseaux ferrés électrifiés (RFE), des réseaux à haute tension (RHT) et loin de ces ouvrages (Ni RFE ni RHT) en considérant les CM sur 24 heures hors sommeil, à l'échelle log à base 10.

4.3.6.2 Les adultes

Sur l'ensemble de la population adulte, 81 habitent à côté des réseaux ferrés électrifiés, 39 ont leurs foyers proches des réseaux à haute tension. Deux personnes habitent à proximité de réseaux à haute tension et de réseaux ferrés électrifiés. Pour réaliser les tests d'égalité des dispersions, nous prenons $h = 6$.

En considérant les moyennes au domicile (MA et MG) avec la période de sommeil ou hors la période de sommeil, les tests donnent les mêmes conclusions. Il n'y a pas de différences entre les moyennes enregistrées dans les foyers proches des réseaux à haute tension et ceux qui sont à côté des réseaux ferrés électrifiés pour les deux types de moyennes. Par contre l'exposition est moins élevée dans les foyers éloignés de ces ouvrages par rapport à ces derniers. On souligne que les tests d'égalité des dispersion ont rejeté cette hypothèse pour les MG. Le test ainsi utilisé pour les paramètres de localisation est celui de Fligner-Policello.

Pour les moyennes sur 24 heures (MA et MG), les résultats des tests restent les mêmes en incluant ou non les CM relatifs à la période de sommeil pour les deux types de moyennes. Il n'y a pas de différences sur les moyennes observées sur les adultes habitant proches des réseaux à haute tension et des réseaux ferrés électrifiés. Elles sont, par contre, moins élevées pour ceux qui habitent dans des foyers isolés de ces ouvrages par rapport à ceux qui ont leurs foyers de résidence à proximité de ces derniers.

4.4 Exposition selon l'activité ou le lieu

Nous sommes tous exposés à un CM résultant de composantes multiples tant à la maison, dans la rue, au travail ou encore dans les transports. Cette exposition dépend non seulement de la présence ou non de toutes sortes de lignes électriques mais aussi des activités que nous exerçons. On peut ainsi se demander quelle est l'exposition pour un type d'activité. Pour répondre à cette question, on se sert des emplois du temps de chaque volontaire et nous calculons les MA et les MG associées à chaque activité ou lieu d'activités. À partir du théorème de la loi des grands nombres, les moyennes associées à chaque activité ou lieu seront estimées. Pour donner des intervalles de confiance, on se servira du théorème central limite.

Théorème 4.2 *: Théorème de la loi faible des grands nombres*
Soient A_1, A_2, $\cdots$, A_n une suite d'expériences aléatoires identiques, et indépendantes les unes des autres. A chaque expérience aléatoire A_i, est associée à une variable aléatoire X_i. On suppose que les variables X_i ont la même espérance mathématique, notée $\mathbb{E}(X) = \mu$, et la même variance notée

σ^2 finies. Soit $\bar{X}_n$, la variable aléatoire définie pour tout entier naturel $n \geq 1$ par :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{4.21}$$

Le théorème dit de la loi faible des grands nombres indique que :

$$\forall \epsilon > 0, \lim_{n \to +\infty} P\left(\left|\bar{X}_n - \mu\right| \leq \epsilon\right) = 1 \tag{4.22}$$

La preuve de ce théorème est donnée en annexe « Complement du chapitre 4 ».
Une autre manière d'écrire le théorème est donnée en (4.23)

$$\forall \epsilon > 0, \lim_{n \to +\infty} P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) = 0 \tag{4.23}$$

Le théorème indique que la moyenne arithmétique des X_i converge en probabilité vers l'espérance commune $\mu = \mathbb{E}(X_i)$ lorsque n tend vers l'infini. Il indique qu'en prenant la moyenne d'un grand nombre de réalisations aléatoires, on peut estimer l'espérance μ avec un fort niveau de certitude.

Dans le cas ici présent, il suffit d'avoir un grand nombre d'individus (une trentaine) pour une activité donnée et l'exposition moyenne associée peut être estimée par la moyenne des expositions observées par les différents individus cette cette activité.

4.4.1 Théorème central limite

La situation est la même que dans le paragraphe précédent : les X_i sont des variables aléatoires indépendantes de même loi. On note μ et σ^2 la moyenne et la variance des X_i. On suppose que σ^2 est finie.

Le théorème précédant indique que $\bar{X}_n$ converge en probabilité vers μ. Si on s'intéresse à la vitesse de convergence, on cherche une équivalence de la suite $\bar{X}_n - \mu$. On est amené à étudier la limite éventuelle de la suite $n^\beta(\bar{X}_n - \mu)$ pour différentes valeurs de β.
Si β est "petit", cette suite va encore tendre vers 0. Elle va diverger si β est « grand ».
On peut espérer que pour une et une seule valeur de β, cette suite converge vers une limite qui n'est ni infinie ni nulle.

Il se trouve que la réponse à cette question a un aspect "négatif" c'est-à-dire que la suite $n^\beta(\bar{X}_n - \mu)$ ne converge même pas en probabilité. Elle a aussi un aspect « positif », cette suite converge, au sens de la convergence en loi, pour la même valeur $\beta = 1/2$ quelque soit la loi des X_i, et toujours vers

une loi normale si $\sigma > 0$. Ce résultat montre pourquoi la loi normale joue un rôle aussi important en probabilité. Il fait l'objet du théorème suivant, appelé théorème central limite ou de la limite centrale.

Théorème 4.3 *Si les X_i sont des variables aléatoires réelles indépendantes et de même loi, appartenant toutes dans L^2, et de moyenne et variance μ et σ^2 ($\sigma > 0$). Alors, la suite de variables aléatoires $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converge en loi vers une variable aléatoire de loi $N(0,1)$. On dit que Z_n converge vers la loi normale centrée et réduite. Une autre manière de le dire est :*

$$\forall t \in \mathbb{R}, \lim_{n \to +\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq t\right) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \tag{4.24}$$

4.4.2 Intervalles de confiance

Pour déterminer un intervalle de confiance pour la moyenne relative à une activité ou un lieu, nous utilisons le théorème central limite ou l'égalité (4.24) qui donne la loi de la limite de Z_n.
Pour établir un intervalle de confiance de μ, il va falloir estimer la variance σ^2. L'estimateur choisi est celui qui est non biaisé (ce n'est pas celui du maximum de vraisemblance). Il est donné par (4.24).

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{4.25}$$

Pour n assez grand, la loi de $\frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}}$ est une loi de Student à $n-1$ degrés de liberté. Un intervalle de confiance de μ de niveau $1-\alpha$ est donné par (4.24) :

$$IC_{1-\alpha}(\mu) = \left[\bar{X}_n - \frac{\hat{\sigma}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \bar{X}_n + \frac{\hat{\sigma}}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right] \tag{4.26}$$

où $t_{1-\frac{\alpha}{2}}(n-1)$ est le quantile à $1-\frac{\alpha}{2}$ d'une loi de Student à $n-1$ degrés de liberté.

La formule (4.26) indique tout simplement que si on pouvait répéter l'expérience une infinité de fois, 95 intervalles sur 100 contiendraient la vraie valeur de μ. On se contente de dire qu'il y a 95% de chance que μ appartienne à $IC_{1-\alpha}(\mu)$.

4.4.3 Estimation de l'exposition par lieu d'activités

Pour quantifier l'exposition par lieu, nous appliquons 4.23 pour estimer la moyenne et 4.26 pour établir un intervalle de confiance de niveau $1-\alpha$.

Du fait que les lieux et les types d'activité sont variés, il n'est pas possible de caractériser l'exposition enregistrée sur tous les lieux et les types d'activité à cause du nombre moins élevé de personnes par lieu. Pour cela, l'estimation des expositions moyennes est basée sur les lieux les plus observés faute de quoi les hypothèses de convergence ne sont pas satisfaites. Pour pouvoir appliquer les égalités (4.23) et (4.26), il faut avoir au moins une trentaine d'observations pour assurer la convergence asymptotique. Cette estimation moyenne permet d'avoir une idée sur l'exposition à l'intérieur de chacun de ces types de lieux. Seul le cas des moyennes arithmétiques est présenté ici. Les résultats sont donnés dans les tableaux 4.3 et 4.4 pour les CM sur 24 heures et dans les tableaux 4.5 et 4.6 pour les moyennes au domicile hors période de sommeil.

Activité	N^*	Temps moyen	Moyenne	IC à 95%
Foyers à côté des RFE	81	7h11min31s	0,10	0,07 - 0,14
Foyers proches des RHT	40	6h29min54s	0,10	0,06 - 0,14
Foyers éloignés des RFE et des RHT	862	7h12min27s	0,07	0,05 - 0,08
Centres commerciaux	84	1h03min22s	0,12	0,08 - 0,16
Trajet en voiture ou en bus	676	0h58min26s	0,08	0,06 - 0,09
Dans les rues	418	0h52min51s	0,07	0,06 - 0,09
École	809	7h15min13s	0,028	0,02 - 0,03

TAB. 4.3 – Estimation des expositions moyennes en μT et des intervalles de confiance pour les MA selon les lieux d'activités pour les enfants. Au domicile les MA sont calculées hors la période de sommeil. Pour les transports ferroviaires, 13 enfants les ont empruntés (trop petit pour généraliser les résultats). RFE=Réseaux ferrés électrifiés, RHT=Réseaux à haute tension et N^* est le nombre de personnes pour l'activité considérée.

Ces résultats montrent que l'exposition est très faible voire négligeable dans les établissements scolaires (tableau 4.3). Les expositions les plus élevées ont été observées dans les transports ferroviaires ou électriques, dans les centres commerciaux et dans les foyers se trouvant à côté des réseaux à haute tension ou des réseaux ferrés électrifiés (tableaux 4.5 et 4.4). Soulignons tout de même les dissimilarités observées sur les deux tableaux pour les expositions au domicile ainsi que la faible représentation d'individus pour certains lieux. Pour un même lieu, il se pourrait que les enfants et les adultes aient des activités différentes. Cela explique pourquoi on a des légères différences sur les moyennes au domicile pour les deux populations. Deux adultes ont enregistré des expositions très élevées dans les transports ferroviaires. Ils ont des moyennes arithmétiques de 35,67 et 10,04 μT. Ils sont retirés lors de l'estimation de l'exposition moyenne dans les transports ferroviaires (ta-

bleau 4.3). Un enfant a aussi enregistré une MA de 8,90 μT dans un centre commercial, il a été retiré lors de l'estimation de l'exposition pour l'activité « Centres commerciaux ».

Activité	N^*	Temps moyen	Moyenne	IC à 95%
Transports ferroviaires	56	00h19min57s	0.46	0,30 - 0,61
Centres commerciaux	338	01h18min47s	0,14	0,11 - 0,17
Foyers proches des RHT	39	10h01min13s	0,11	0,06 - 0,17
Foyers à côté des RFE	81	10h22min57s	0,09	0,07 - 0,12
Foyers éloignés des RHT et des RFE	936	09h25min32s	0,08	0,06 - 0,10
Transport en voiture ou bus	794	01h45min01s	0,14	0,09 - 0,19
Travail sur ordinateur au bureau	342	06h51min52s	0,10	0,08 - 0,12
Dans les rues	314	01h23min29s	0,10	0,08 - 0,12

TAB. 4.4 – Estimation des expositions moyennes en μT et des intervalles de confiance à 95% pour les MA selon les lieux d'activités hors la période de sommeil pour les adultes.

En considérant les CM sur 24 heures, des différences apparaissent en termes d'exposition entre les moyennes au domicile pour les enfants et celles des adultes (tableaux 4.5 et 4.6). Ces différences peuvent être expliquées par la combinaison des diverses activités que font les adultes (ménage, repassage, cuisine, ...) et les radio-réveils.

Lieu	N^*	Temps Moyen	Moyenne	IC à 95%
Foyers proches des RHT	40	16h50min25s	0,10	0,04 - 0,15
Foyers à côté des RFE	81	17h11min48s	0,16	0,02 - 0,29
Foyers éloignés des RHT et des RFE	862	17h20min12s	0,11	0,08 - 0,14

TAB. 4.5 – Estimation des expositions moyennes en μT et des intervalles de confiance pour les MA observés aux domiciles pour les enfants en incluant la période de sommeil.

Lieu	N^*	Temps moyen	Moyenne	IC à 95%
Foyers à côté des RHT	39	18h26min41s	0,31	0,07 - 0,56
Foyers proches des RFE	78	17h58min53s	0,09	0,07 - 0,11
Foyers éloignés des RHT et des RFE	936	17h53min43s	0,14	0,11 - 0,18

TAB. 4.6 – Estimation des expositions moyennes en μT et des intervalles de confiance pour les MA observés aux domiciles pour les adultes en incluant la période de sommeil. Trois personnes habitant dans des foyers proches des RFE (réseaux ferrés électrifiés) ont observé au domicile une moyenne de 4,54 μT. Ils sont retirés lors de l'estimation des moyennes de ce tableau.

4.5 Conclusion

Dans ce chapitre, les expositions moyennes ont été analysées selon deux scénarios (en considérant les CM sur 24 heures et ceux enregistrés hors période de sommeil). Les résultats dépendent largement du scénario considéré. Les MA et MG observées sur 24 heures sont respectivement de 0,09 et 0,02 μT pour les enfants et 0,14 et 0,03 μT pour les adultes. Au total, 30 enfants ($3,1\%$) ont observé une MA supérieure à 0,4 μT. Deux d'entre eux ont observé à une MG supérieure à cette valeur. Les sources liées à ces valeurs élevées sont à 80% des radio-réveils. Le même constat est aussi observé chez les adultes les plus exposés, $81,8\%$ des MA supérieures au quantile à $99,0\%$ sont associées à des CM générés par des radioréveils. Pour ces sources, nous ne pouvons pas justifier avec certitude que les CM générés par ces dernières et enregistrés par les EMDEX la nuit reflètent l'exposition des personnes ou non. C'est pourquoi nous disons que $3,1\%$ des enfants ont *observé* une MA supérieure à 0,4 μT au lieu d'utiliser le mot *exposé*. Cette proportion est plus élevée que celle attendue, en comparaison avec les autres pays.

En considérant les CM enregistrés hors la période de sommeil, nous observons 11 enfants exposés à une MA supérieure à 0,4 μT soit $1,1\%$ des enfants (cette proportion est pratiquement trois fois plus élevée que celle observée dans l'étude UKCCS [13] et est légèrement supérieure à celle observée dans l'étude d'Ahlbom [14]). Une explication possible est le fait que l'exposition dans l'étude UKCCS est une mesure en point fixe et non une mesure personnelle. Les expositions moyennes deviennent 0,05 et 0,02 μT respectivement pour MA et MG pour les enfants et 0,10 et 0,03 μT pour les adultes. Ces résultats montrent que les mesures sur 24 heures surestiment l'exposition à cause des CM émis par les radio-réveils.

Les études de comparaison des moyennes ont montré que globalement les

enfants sont moins exposés que les adultes, et que les personnes sont plus exposées en Ile-de-France que dans les autres régions. Elles ont surtout montré que l'exposition des personnes ne peut pas se résumer au CM enregistré au domicile tout simplement parce que les enfants sont plus exposés au domicile qu'à l'extérieur alors que pour les adultes, c'est le contraire. D'autres tests basés sur les CM enregistrés au domicile ont été réalisés comme la comparaisons des moyennes jour/ nuit ou encore la comparaison des moyennes dans les foyers proches des réseaux à haute tension, des réseaux ferrés électrifiés et dans ceux éloignés de ces ouvrages. Les résultats de ces tests permettent de conclure qu'il n'y a pas de différence entre les moyennes observées dans les foyers proches des lignes à haute tension et dans ceux qui sont à côté des réseaux ferrés électrifiés (sauf pour les enfants en MG). Ces résultats ont été aussi retrouvés en considérant les moyennes sur 24 heures. Dans les foyers les plus éloignés, l'exposition est moins élevée que dans ces derniers.

Au domicile, les enfants ont des moyennes globalement plus importantes le jour que la nuit. Pour les adultes, la conclusion dépend des moyennes considérées. Les MA sont plus élevées le jour que la nuit alors que les MG varient dans le sens contraire. Ces derniers résultats sont à prendre avec précaution car les écarts des fonctions de répartition sont loin d'être constants.

Ces tests nous ont conduit à estimer l'exposition moyenne observée dans les différents lieux. Cette quantification a montré que les moyennes les plus élevées ont été enregistrées dans les transports ferroviaires (MA=0,46 ; IC=[0,30 ; 0,61]) et dans les centres commerciaux (MA=0,14 ; IC=[0,11 ; 0,17]) pour les adultes. Dans les établissements scolaires, l'exposition est très faible (MA=0,03 ; IC=[0,02 ; 0,03]). Au domicile, les moyennes observées dépendent du scénario considéré. Il est par contre difficile de conclure que les CM enregistrés aux domiciles représentent réellement l'exposition de certaines personnes ayant posé l'appareil à côté du radio-réveil. En effet, les valeurs sont alors plus élevées et d'une part, ne représentent pas l'exposition de la personne et d'autre part masquent la présence éventuelle d'autres sources de champ magnétique, le champ mesuré étant la somme des champs venant de toutes les sources. Cela peut induire une surestimation de l'exposition au domicile et donc sur 24 heures. Éliminer les CM relatifs à la période de sommeil conduit à une perte d'information mais permet de s'assurer que les CM analysés reflètent l'exposition des personnes.

Le fait de savoir si un foyer est à proximité d'ouvrages électriques (lignes à haute tension ou réseaux ferrés électrifiés), ou bien le fait de connaître le temps passé dans les transports ferroviaires peut-il permettre de prédire l'exposition moyenne ou une partie de l'exposition moyenne ? Pour cela, nous étudions dans le chapitre suivant les variables qui caractérisent l'exposition.

Chapitre 5

Caractérisation des expositions moyennes

5.1 Introduction

En première année de thèse, nous avons réalisé une analyse sur les dépendances entre les expositions moyennes et des variables explicatives des données de la première phase. Elle a permis d'identifier des corrélations entre les moyennes et certaines variables. Pour caractériser ces structures, nous avons *a priori* choisi des modèles linéaires. Ces modèles ont donné des taux de variance expliquée très faibles. Le taux de la variance expliquée R^2 d'un modèle est défini par :

$$R^2 = 1 - \frac{\tilde{S}_m^2}{\tilde{S}_0^2}$$

où $\tilde{S}_m^2$ est l'estimateur non biaisé de la variance des résidus du modèle et $\tilde{S}_0^2$ celui du modèle constant.

Ce problème peut être expliqué de trois manières :

1. Les relations détectées sont linéaires mais les informations dont on dispose ne permettent pas à elles seules de bien caractériser les expositions moyennes. Dans ce cas il faudrait introduire de nouveaux facteurs pouvant influencer les expositions pour espérer améliorer les taux des variances expliquées.
2. Les relations identifiées ne sont pas linéaires. Le fait d'avoir choisi a priori un modèle linéaire induit une perte d'information. Pour apporter des améliorations, il faut trouver les bonnes relations entre les variables explicatives continues et la variable dépendante. Une manière de faire est de ne pas supposer de relations particulières mais de modéliser ces structures de relation comme étant des fonctions inconnues qu'on peut estimer à l'aide de méthodes non paramétriques.
3. Les informations à disposition sont insuffisantes pour caractériser les expositions et les relations ne sont pas linéaires.

Le modèle linéaire repose sur un postulat qui doit être vérifié pour être valide. La plupart des méthodes développées au début de l'apparition des statistiques et encore utilisées de nos jours font appel à des hypothèses qui sont parfois très contraignantes. Elles restreignent considérablement l'étendue des applications des dites méthodes. L'augmentation de la puissance de calculs des ordinateurs et les recherches ont permis d'assouplir certains de ces postulats et ainsi d'obtenir des modèles flexibles susceptibles de représenter la réalité. Les méthodes de régression non paramétrique en sont un bon exemple.

Lorsqu'on veut étudier la relation entre une variable dépendante Y et une variable explicative X, on peut utiliser la régression linéaire. Cette méthode

est très pratique lorsqu'elle est appropriée car elle suppose un modèle simple. Ce modèle est donné par (5.1).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n \tag{5.1}$$

On suppose que les ε_i sont des variables aléatoires indépendantes et identiquement distribuées de loi normale centrée et de variance σ^2.

Lorsque ce postulat est vérifié, on peut réaliser des tests sur les paramètres. Ce modèle impose un postulat très restrictif sur la linéarité de la liaison entre les deux variables. Ce postulat n'est pas toujours respecté. Pour expliquer la relation entre Y et X, on peut spécifier une autre forme de liaison ou encore transformer les variables et appliquer un modèle linéaire. Toutefois, il n'est pas évident de trouver la transformation appropriée. Dans ces conditions, déduire la bonne relation entre Y et X devient rapidement complexe. De nouvelles méthodes de régression plus souples permettant de lever l'a priori de la relation entre les variables et s'adaptant à une large classe de données sont développées. Elles sont appelées des méthodes de *régression non paramétrique*.

Ces méthodes peuvent aussi servir à décrire la forme que devrait prendre un modèle de régression paramétrique. Parmi elles, les méthodes à noyau, les fonctions splines (les splines de régression et les splines de lissage), etc. Toutes ces méthodes permettent de contrôler la flexibilité de l'estimateur. Cette flexibilité a un prix, et comme dans le cas paramétrique, elles doivent composer avec la dualité biais-variance. Ainsi, le fait de suivre fidèlement les données augmente la variance de l'estimateur, alors qu'un estimateur plus lisse augmente le biais. Il faut trouver un compromis entre le biais et la variance de l'estimateur.

On peut aussi s'intéresser aux effets simultanés de plusieurs variables sur une variable réponse. La solution qui apparaît de manière naturelle est d'appliquer une régression linéaire multiple : la généralisation de (5.1). Le principal défaut de cette généralisation est la linéarité de la relation. Cette hypothèse implique alors que la forme de la relation est un hyperplan dans un espace dont la dimension dépend du nombre de variables impliquées dans la relation. Lorsque cette hypothèse n'est pas satisfaite, il devient difficile d'imaginer la relation appropriée. Une des solutions proposées dans la littérature est la version multivariée de la régression non paramétrique [41].

Dans le cadre de notre application, pour espérer mieux caractériser les expositions moyennes, nous modélisons les relations entre les variables continues et la variable réponse par des fonctions qu'on estime à l'aide de méthodes non paramétriques. Des tests de comparaison peuvent être réalisés entre

le modèle linéaire multiple et le modèle non paramétrique afin de voir l'amélioration apportée par le modèle non paramétrique.

5.2 Régression non paramétrique univariée

Afin de mieux comprendre les méthodes de régression utilisées dans le cas multivarié, nous présentons dans cette section les concepts sous-jacents à ces méthodes dans le cadre d'une régression univariée.

Les méthodes de régression univariée sont généralement utilisées pour décrire ou modéliser la relation entre une variable dépendante Y et une variable explicative X, sans supposer une forme *a priori* ou particulière.
Il existe plusieurs manières pour estimer une fonction de façon non paramétrique. Dans cette section, nous présentons deux méthodes connues sous les noms de méthode *loess* (l'une des méthodes les plus connues de la littérature) et *splines de lissage*. Dans le cas multivarié, c'est la généralisation de la seconde méthode qui sera traitée car elle permet de réaliser des tests de sous modèles.

5.2.1 Généralités sur les fonctions de lissage

Soit $(x_i, y_i)_{i=1,\ldots,n}$, un échantillon aléatoire d'une variable (X, Y) où les x_i représentent les valeurs observées de la variable explicative X et les y_i celles de la variable dépendante Y. La relation entre les x_i et les y_i est modélisée par (5.2).

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, n \tag{5.2}$$

où les ε_i sont des variables non corrélées entre elles de moyenne nulle et de variance σ^2. f est la fonction de régression que l'on cherche à estimer. Pour pouvoir faire des tests sur l'estimateur, il est nécessaire de faire des hypothèses sur les ε_i. On fait souvent l'hypothèse d'un bruit blanc gaussien.

Les estimateurs de f obtenus de façon non paramétrique sont généralement appelés fonctions de lissage. Ce nom vient du fait qu'elles lissent les données de l'échantillon pour obtenir des estimateurs. Ce lissage compose avec le biais et la variance de l'estimateur.

5.2.1.1 Le compromis biais-variance

Le compromis entre le lissage et la flexibilité de l'estimateur est identifié comme la dualité biais-variance. En augmentant la flexibilité, il est possible de suivre plus fidèlement les données, on réduit ainsi le biais de l'estimateur. L'estimateur ainsi obtenu aura tendance à osciller, ce qui fait augmenter la variance. L'idéal est d'avoir une courbe qui soit assez lisse avec moins

de variance. Pour ce faire, il faut diminuer la flexibilité de l'estimateur, ce qui implique de suivre moins fidèlement les données, donc d'augmenter le biais. Ainsi tout utilisateur d'un modèle de régression non paramétrique doit composer avec cette dualité au moment du choix du paramètre de lissage noté λ.

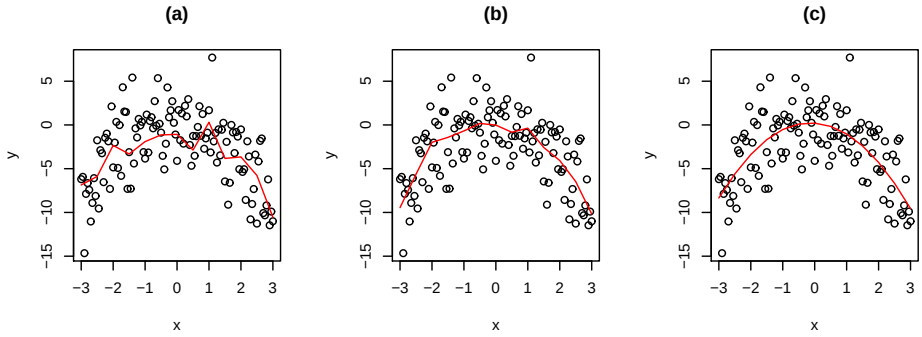

FIG. 5.1 – Illustration du compromis entre biais et variance. La fonction à estimer est $f(x) = -x^2$ à laquelle nous avons rajouté un bruit gaussien de moyenne 0 et de variance 3, $x \in [-3, 3]$.

La figure 5.1 illustre cette dualité. Elle montre que la courbe (a) représentant une fonction de lissage *loess* avec un paramètre de lissage fixé à $\lambda = 0,1$ est plus variable que les deux autres pour lesquelles ce paramètre est fixé à $\lambda = 0,3$ et $\lambda = 2$. L'utilisation d'un paramètre de lissage plus petit permet d'avoir une fonction qui suit bien les données (moins de biais) mais avec plus de variance. Inversement, prendre une grande valeur pour ce paramètre permet d'avoir un estimateur plus lisse avec moins de variance mais avec plus de biais. Le compromis entre le biais et la variance est contrôlé par ce paramètre appelé aussi pénalité.

5.2.1.2 Matrice de lissage et degrés de liberté

Dans la plupart des cas, les fonctions de lissage sont obtenues par une combinaison linéaire des observations (5.3).

$$\hat{f} = \mathbf{S}_\lambda \mathbf{y}. \tag{5.3}$$

où $\mathbf{S}_\lambda$ est la matrice de lissage et $\mathbf{y}$ le vecteur composé des y_i. $\hat{f}$ est alors un vecteur de dimension n. La détermination de la matrice de lissage dépend de la méthode utilisée, du paramètre de lissage ainsi que de la façon dont les x_i sont distribués. La matrice $\mathbf{S}_\lambda$ est très utile pour la détermination du nombre de degrés de liberté (*ddl*) d'une fonction de lissage. En régression paramétrique, plus le nombre de degrés de liberté est élevé, meilleur est l'ajustement. Cette affirmation n'est plus valable dans le cas non paramétrique

car les modèles ne sont pas exprimés en termes de paramètres (mis à part le paramètre de lissage). Pour avoir une mesure permettant de comparer la flexibilité de deux estimateurs de f, on utilise le même concept que dans le cas de la régression paramétrique.

En régression linéaire (5.1), les degrés de liberté d'un modèle (nombre de paramètres) peuvent être obtenus en calculant la trace de la matrice chapeau [38]. Dans le cas d'une régression linéaire, cette matrice est définie par $\mathbf{H}_0 = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ où $\mathbf{X}$ est la matrice dont la première colonne est composée de 1 et les autres colonnes sont les vecteurs des variables explicatives. Dans ces conditions, le vecteur prédit $\hat{\mathbf{y}}$ est donné par $\hat{\mathbf{y}} = \mathbf{H}_0\mathbf{y}$. En faisant le parallèle avec la matrice de lissage, le nombre de *ddl* peut être calculé comme (5.4).

$$ddl = \mathrm{tr}(\mathbf{S}_\lambda) \tag{5.4}$$

Certains auteurs utilisent une autre définition pour estimer les degrés de liberté en prenant la trace de la matrice $\mathbf{S}_\lambda\mathbf{S}_\lambda^\top$. De façon générale, plus le nombre de degrés de liberté associé à une fonction de lissage est grand, plus cette dernière est flexible. Certaines méthodes permettent de fixer le nombre de degrés de liberté souhaité au lieu du paramètre de lissage.

Lorsqu'on cherche à faire des tests pour comparer différents modèles, on a plutôt besoin d'une mesure du nombre de degrés de liberté pour la distribution des erreurs. Dans ce cas, une autre définition est utilisée. Elle est donnée en (5.5) par Hastie et Tibshirani [39].

$$ddl_{\text{erreur}} = n - \mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda^\top) \tag{5.5}$$

5.2.1.3 Critères de sélection du paramètre de lissage

En général, les utilisateurs de fonction de lissage cherchent à obtenir l'ajustement optimal pour la relation qu'ils veulent estimer. On doit donc trouver le paramètre de lissage qui permet d'avoir le meilleur compromis entre le lissage et la flexibilité (le biais et la variance). On cherche donc à obtenir l'estimateur $\hat{f}_\lambda$ qui se rapproche le plus possible de la vraie fonction f de l'équation (5.2). La meilleure façon de mesurer la précision d'un estimateur serait d'utiliser un échantillon complémentaire formé de nouvelles observations et de minimiser les erreurs de prédiction. On choisirait donc l'estimateur $\hat{f}_\lambda$ qui minimise les erreurs quadratiques de prédiction. Ce critère est connu sous le nom de PSE (Average Predictive Squared Error) et est donné par (5.6).

$$PSE(\lambda) = \frac{1}{\breve{n}} \sum_{i=1}^{\breve{n}} \mathbb{E}[\breve{y}_i - \hat{f}_\lambda(\breve{x}_i)]^2 \tag{5.6}$$

où $(\breve{x}_i, \breve{y}_i)_{i=1,\ldots,\breve{n}}$ est un nouveau échantillon de (X, Y) ce qui implique que $\breve{y}_i = f(\breve{x}_i) + \breve{\varepsilon}_i$ où $\breve{\varepsilon}_i$ est non corrélée avec les autres erreurs ε_i [39]. On obtient ainsi une méthode de sélection automatique du paramètre de lissage.
Il existe plusieurs moyens permettant d'estimer le PSE défini en (5.6). La méthode la plus utilisée est celle de la validation croisée. Elle est donnée par :

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\lambda^{-i}(x_i) \right)^2 \tag{5.7}$$

où $\hat{f}_\lambda^{-i}(x_i)$ est l'estimateur obtenu au point x_i lorsque l'observation (x_i, y_i) est exclue de l'échantillon pour le calcul de l'estimation.
On calcule la valeur du critère pour une série de valeurs de paramètres de lissage puis on choisit celui qui minimise le critère. Le critère (5.7) est la version *leave one out* de la validation croisée. Pour optimiser le temps de calcul, on utilise une approximation du critère de validation croisée généralisée GCV (Generalized Cross-Validation) appelée aussi somme pénalisée des résidus au carré. Elle est donnée par :

$$GCV(\lambda) = \frac{\sum_{i=1}^{n} \left(y_i - \hat{f}_\lambda(x_i) \right)^2}{\left(1 - \frac{c(\lambda)}{n} \right)^2} \tag{5.8}$$

où $c(\lambda)$ est une fonction du paramètre de lissage λ. Elle varie selon les auteurs et le type de régression non paramétrique utilisé. Dans le cas des fonctions de lissage linéaires, Hastie et Tibshirani proposent de prendre
$c(\lambda) = \mathrm{tr}(\mathbf{S}_\lambda)$ [39].

Remarque
La plupart des méthodes ont été développées en supposant que la variable prédictive X est continue. Les x_i sont ainsi supposés différents deux à deux c'est-à-dire que $x_i \neq x_j$ pour $i \neq j$.
Dans la réalité, il n'est pas rare d'observer des égalités dans la variable prédictive. Il existe un moyen simple pour traiter ces égalités. Ce dernier consiste à considérer la moyenne de la variable réponse pour tous les points ayant la même valeur x_i et de lui associer un poids w_i égal au nombre d'égalités pour l'observation x_i.
On obtient ainsi un échantillon $(x_i, moyenne_{x_i}(y_i), w_i)$,$i = 1, \ldots, n^*$. Où n^* est le nombre de valeurs distinctes dans l'échantillon $(x_i)_{i=1, \ldots, n}$. Dans ces conditions, la régression est appelée régression non paramétrique pondérée.

5.2.1.4 Tests de comparaison des fonctions de lissage

En statistique classique, on fait des tests pour voir si les paramètres estimés sont significatifs. En faisant une similitude, des tests approximatifs permettant de guider la sélection du modèle ont été développés. Ces tests sont construits par analogie avec la régression linéaire [39]. Pour deux estimateurs $\hat{f}_1$ et $\hat{f}_2$ où $\hat{f}_2$ est le plus flexible, supposons que l'on désire tester les hypothèses suivantes :

H_0 : $\hat{f}_1$ et $\hat{f}_2$ sont équivalents
contre
H_1 : $\hat{f}_1$ et $\hat{f}_2$ sont différents

On utilise la statistique de Fisher définie par :

$$F = \frac{(RSS(\hat{f}_1) - RSS(\hat{f}_2))/(\gamma_1 - \gamma_2)}{RSS(\hat{f}_2)/\gamma_2} \tag{5.9}$$

qui suit, sous H_0, une loi de Fisher à $(\gamma_1 - \gamma_2, \gamma_2)$ degrés de liberté.
RSS est la somme des résidus au carré et γ_k est le nombre de degrés de liberté de l'erreur de l'estimateur $\hat{f}_k$ défini par la formule (5.5). On utilise aussi ce test pour comparer différents modèles non paramétriques utilisant les mêmes données. C'est ce test qui sera utilisé pour tester la significativité d'une variable dans le cas multivarié.
Il faut toutefois garder en tête que ces tests dépendent du choix de la méthode de lissage utilisée et des paramètres associés.

Dans le cas particulier où le modèle testé est linéaire, le test se resume à un test de non linéarité. Cela revient à tester l'hypothèse :

H_0 : $f(x_i) = \beta_0 + \beta_1 x_i$ (la relation qui lie les deux variables est linéaire)
contre
H_1 : $f(x_i) \neq \beta_0 + \beta_1 x_i$ (la relation n'est pas linéaire)

Pour réaliser ce test, on estime les paramètres β_0 et β_1 du modèle sous H_0 et f sous H_1 puis on calcule la somme des résidus sous chaque hypothèse notée respectivement RSS_0 et RSS_1. La statistique de test est définie par (5.10).

$$F = \frac{(RSS_0 - RSS_1)/(ddl_1 - 2)}{RSS_1/(n - ddl_1)}. \tag{5.10}$$

où ddl_1 est le nombre de degré de liberté du modèle sous l'hypothèse H_1. Il peut être calculé à l'aide de l'expression (5.4).
Sous l'hypothèse nulle (H_0), F suit une loi de Fisher à $(ddl_1 - 2)$ et $(n - ddl_1)$ degrés de liberté. Le test rejette H_0 si la valeur observée de F est supérieure

au quantile à $1-\alpha$ d'une loi de Fisher à $((ddl_1 - 2), (n - ddl_1))$ degrés de liberté.

5.2.2 La méthode loess

La méthode *loess* (Locally weighted running-line, en anglais) a été introduite par Cleveland [40]. Elle est généralement la plus utilisée à cause de sa simplicité et de sa rapidité par rapport autres méthodes. Elle peut facilement être utilisée pour ajuster des modèles multidimensionnels.

5.2.2.1 Forme de l'estimateur

La méthode utilise les moindres carrés pour estimer la fonction f du modèle (5.2). L'estimateur associé obtenu n'est pas représenté par une équation unique mais par un ensemble de points. Soit un échantillon (x_i, y_i), $i = 1, \ldots, n$ où on suppose que les valeurs x_i de la variable explicative sont toutes distinctes. La procédure d'estimation par la méthode *loess* est la suivante :

Méthode loess

1. Pour tout point x_0 du domaine de X, on choisit les k plus proches voisins, appelés voisinage, et on calcule la distance entre ces points et le point x_0. On note $N(x_0)$, l'ensemble de points constituant le voisinage de x_0. Le cardinal de $N(x_0)$ (k) est fixé par le paramètre de lissage λ désignant ici, la proportion des voisinages de x_0 par rapport à n. Cette proportion est maintenue constante tout au long du processus.
2. On donne à chaque point du voisinage de x_0 un poids inversement proportionnel à sa distance par rapport à x_0 à l'aide de la fonction tricubic donnée par (5.11).

$$w(u) = \begin{cases} (1-u^3)^3, & \text{pour } 0 \leq u < 1 \\ 0 & \text{sinon} \end{cases} \tag{5.11}$$

   Pour que le poids soit inversement proportionnel à la distance, on prend

$$u_i = \frac{|x_0 - x_i|}{\max_{N(x_0)} |x_0 - x_i|}. \tag{5.12}$$

   pour chaque point x_i du voisinage $N(x_0)$.
3. On calcule l'estimateur de f au point x_0 en utilisant le polynôme de degré déterminé par l'utilisateur, estimé en appliquant la méthode de moindres carrés pondérés à l'ensemble des points du voisinage $N(x_0)$ [38].

Concrètement, cette démarche n'est pas appliquée à tous les points d'intérêt afin de limiter le temps d'exécution. On choisit un ensemble de points répartis sur le domaine de X pour lesquels la démarche est appliquée puis on utilise une méthode d'interpolation pour obtenir le résultat pour les autres points d'intérêt.
C'est une méthode facile à mettre en œuvre et qui permet de controler à la main la dualité biais-variance. C'est aussi un très bon outil lorsque l'on désire connaître l'allure générale de la courbe dans le but de faire une régression paramétrique. Par contre, puisque c'est une méthode plutôt heuristique, elle ne permet pas à l'utilisateur d'obtenir une expression simple pour la forme de la relation ni d'effectuer des tests exacts comme ceux qui sont faits en régression paramétrique.

La figure 5.2 est un exemple de l'estimateur *loess* pour des polynômes de degré 1 et 2.

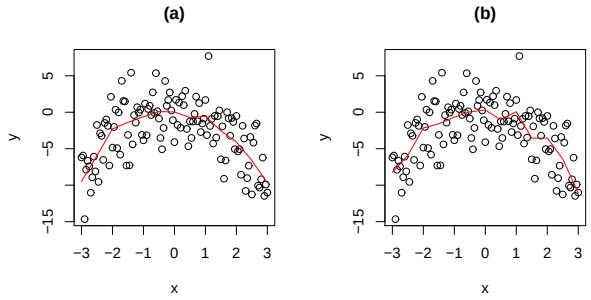

FIG. 5.2 – Illustration des différences observées sur l'estimateur *loess* selon le degré du polynôme. Les fonctions de lissage représentées utilisent des polynômes de degré 1 pour (a) et 2 pour (b). Le paramètre de lissage utilisé est 0,2 pour les deux courbes. La fonction à estimer est $f(x) = -x^2$ à laquelle est rajouté un bruit gaussien de moyenne 0 et de variance 3, $x \in [-3, 3]$

Il existe des différences sur l'estimateur *loess* selon le degré du polynôme d'interpolation choisi. La figure 5.2 illustre ce phénomène. Pour les polynômes de degré 1, l'estimateur ressemble à des fonctions linéaires continues par morceaux alors que pour les polynômes de degré 2, ce dernier ressemble à des fonctions quadratiques continues par morceaux.

La méthode ne spécifie pas le degré du polynôme utilisé pour obtenir l'estimateur des moindres carrés. Les polynômes de degré un ou deux sont souvent utilisés. Ils permettent d'avoir des résultats satisfaisants avec un temps minimal.

5.2.3 Les splines de lissage

Définition 1 *Une fonction spline de degré $m-1$ est une fonction continue, définie par intervalles et dont les morceaux sont des polynômes de degré $m-1$. Elles peuvent s'écrire sous la forme :*

$$s_\Lambda(x) = \sum_{j=1}^{m} \theta_j x^{j-1} + \sum_{j=1}^{k} \delta_j (x - \zeta_j)_+^{2m-1} \tag{5.13}$$

avec

$$u_+ = \begin{cases} u & si\ u \geq 0 \\ 0 & si\ u < 0 \end{cases}$$

L'ensemble de points $\{\zeta_1, ..., \zeta_k\}$, appelé ensemble de nœuds Λ, définit les points des coupures des intervalles. Les paramètres sont les θ_j et les δ_j.

Les splines de lissage sont une façon d'utiliser les fonctions splines pour estimer la fonction de régression du modèle (5.2). Contrairement à d'autres méthodes comme les splines de régression qui utilisent plutôt des méthodes intuitives ou d'essais et erreurs pour déterminer l'ensemble des noeuds Λ (et par conséquent l'estimateur s_Λ), les splines de lissage minimisent un critère bien précis (5.14). Celui-ci combine la mesure classique de la qualité de l'ajustement, la somme des erreurs quadratiques et une mesure de la qualité de lissage sous la forme (5.14).

$$\sum_{i=1}^{n} (y_i - s_\Lambda(x_i))^2 + \lambda \int s_\Lambda^{(m)}(t)^2 dt. \tag{5.14}$$

où λ est le paramètre de lissage prenant ses valeurs dans $[0, +\infty[$ et m est fixé et sert à définir le degré des polynômes ajustés. La forme de l'estimateur assure que les $m-2$ premières dérivées sont continues, ce qui permet d'obtenir une courbe assez lisse, selon la valeur de m. Les splines les plus fréquemment utilisées sont les splines cubiques, qui sont composées de polynômes de degré 3 et dont les deux premières dérivées sont continues, ainsi que les splines linéaires, composées de polynômes de degré un. Dans cette étude, on fixe $m = 2$ et on utilise des pôlynomes cubiques. Plus la valeur de λ est proche de 0, plus l'estimateur est flexible. Par contre, lorsqu'on augmente la valeur de λ, on donne plus d'importance à la deuxième partie du critère (5.14), ce qui diminue l'intégrale et donc rend l'estimateur plus lisse.

5.2.3.1 Forme de l'estimateur

Eubank a montré en 1999 [41] que l'unique fonction parmi l'ensemble des fonctions dont, les dérivées $f^{(0)}$, $f^{(1)}$, ..., $f^{(m-1)}$ sont absolument continues et dont la $m^{\text{ème}}$ dérivée est de carré intégrable, qui minimise le

critère (5.14), est une fonction spline de degré $2m-1$ avec des nœuds à chacune des valeurs distinctes de la variable X dans l'échantillon. On ajoute la contrainte d'être formée de polynômes de degré d en dehors de l'intervalle de couverture de la variable explicative X. Les splines possédant une telle contrainte sont appelées des splines naturelles.

Il est à noter que, même si la formule (5.13) porte à croire que le nombre de paramètres à estimer est $n^* + m - 1$ (n^* est le nombre des x_i dont les valeurs sont distinctes entre elles), les contraintes imposées aux extrémités de chaque intervalle réduisent le nombre de paramètres à n^*. On évite ainsi la surparamétrisation.

On note $(x_i^*)_{i=1,\ ...,\ n^*}$ le sous échantillon de X composé des observations distinctes des $(x_i)_{i=1,\ ...,\ n}$ et rangées par ordre croissant (les x_i^* sont appelés des nœuds). Construire f en utilisant des splines cubiques revient à la représenter en joignant les nœuds par des portions de polynômes cubiques de manière continue de tel sorte que la dérivée seconde soit continue en ces points. En définissant la base $(b_j(x))_{j=1,\ ...,\ n^*+2}$ par (5.15) et en supposant que f est une fonction spline qui peut s'écrire dans cette base par l'expression donnée en (5.16), f est une fonction spline cubique à laquelle on impose que la dérivée seconde soit nulle en dehors du domaine de X c'est à dire $\sum_{j=1}^{n^*} \beta_j = 0$ et $\sum_{j=1}^{n^*} \beta_j x_j^* = 0$.

$$b_j(x) = \begin{cases} \mid x - x_j^* \mid^3, & j = 1,\ \ldots,\ n^* \\ 1 & \text{si } j = n^* + 1 \\ x & \text{si } j = n^* + 2 \end{cases} \tag{5.15}$$

$$f(x) = \sum_{j=1}^{n^*+2} \beta_j b_j(x) \tag{5.16}$$

En supposant que la fonction f reliant les x_i et les y_i dans (5.1) est une fonction spline, on peut l'écrire dans la base $(b_j(x))_{j=1,\ ...,\ n^*}$ par (5.17) où le vecteur β composé des β_j, $1 \leq j \leq n^*$ est le vecteur des paramètres qu'on souhaite estimer.

$$f(x) = \sum_{j=1}^{n^*} \beta_j b_j(x) \tag{5.17}$$

Pour estimer f tout en contrôlant la flexibilité par rapport aux données (5.14), on calcule f' puis f'' par :

$f'(x) = \sum_{j=1}^{n^*} \beta_j b_j'(x)$ puis $f''(x) = \sum_{j=1}^{n^*} \beta_j b_j''(x)$

f'' peut s'écrire comme l'expression (5.18).

$$f''(x) = \sum_{j=1}^{n^*} \beta_j b_j''(x) = \tilde{\mathbf{b}}''(x)^\top \beta \tag{5.18}$$

où $\tilde{\mathbf{b}}''(x)$ est le vecteur des dérivées secondes de la base prises en x.
Ainsi $[f''(x)]^2$ est une forme quadratique. Sa formule est donnée en (5.19).

$$[f''(x)]^2 = \beta^\top \tilde{\mathbf{b}}''(x)^\top \tilde{\mathbf{b}}''(x)\beta = \beta^\top \mathbf{G}(x)\beta \tag{5.19}$$

où $\mathbf{G}(x)$ est la matrice définie par :

$$\mathbf{G}(x) = \begin{pmatrix} b_1''(x)^2 & b_1''(x)b_2''(x) & b_1''(x)b_3''(x) & \dots & b_1''(x)b_{n^*}''(x) \\ b_2''(x)b_1''(x) & b_2''(x)^2 & b_2''(x)b_3''(x) & \dots & b_2''(x)b_{n^*}''(x) \\ b_3''(x)b_1''(x) & b_3''(x)b_2''(x) & b_3''(x)^2 & \dots & b_3''(x)b_{n^*}''(x) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{n^*}''(x)b_1''(x) & b_{n^*}''(x)b_2''(x) & b_{n^*}''(x)b_3'(x) & \dots & b_{n^*}''(x)^2 \end{pmatrix}$$

Ainsi $J(f) = \int f''(x)^2$ peut se calculer à partir de (5.20).

$$J(f) = \beta^\top \int \mathbf{G}(x)dx\beta = \beta^\top \mathbf{S}(x)\ \beta \tag{5.20}$$

Pour une base donnée, on peut toujours calculer les coefficients de la matrice $\mathbf{S}$ et ensuite écrire le terme de pénalité $J(f)$ comme une forme quadratique des paramètres du vecteur β (la matrice $\mathbf{S}$ d'ordre $n^* \times n^*$ ne dépend pas de β).

Pour estimer la fonction de lissage (5.2) dans une base $(b_j(x))_{j=1,\ ...,\ n^*}$ donnée en utilisant le critère (5.14), on définit :

- $\mathbf{X} = (b_j(x_i^*))_{i,j=1,\ ...,\ n^*}$ la matrice des bases des splines naturelles.
- $\mathbf{W}$ la matrice diagonale des poids (si tous les x_i sont distinctes alors $\mathbf{W}$ est la matrice identité).
- $\mathbf{y}^*$, le vecteur composé des moyennes des y_i par rapport au nombre de répétitions des x_i^*.
- On calcule les coefficients de la matrice $\mathbf{S}$ comme dans (5.20).

Pour un paramètre de lissage λ donné, estimer f en minimisant (5.14) dans une base donnée revient à estimer les paramètres du vecteur β minimisant (5.21).

$$\|\mathbf{W}^{1/2}(\mathbf{X}\beta - \mathbf{y}^*)\|^2 + \lambda\beta^\top \mathbf{S}\beta \tag{5.21}$$

En développant (5.21), on a :

$$(\mathbf{X}\beta - \mathbf{y}^*)^\top \mathbf{W}(\mathbf{X}\beta - \mathbf{y}^*) + \lambda\beta^\top \mathbf{S}\beta = \beta^\top(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda\mathbf{S})\beta + 2\beta^\top \mathbf{X}\mathbf{W}\mathbf{y}^* + \mathbf{y}^{*\top}\mathbf{W}\mathbf{y}^*$$

$$(\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda\mathbf{S})\beta = \mathbf{X}^\top \mathbf{W}\mathbf{y}^* \tag{5.22}$$

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{W}\mathbf{X} + \lambda \mathbf{S})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y}^* \tag{5.23}$$

Le paramètre de lissage λ est estimé en utilisant le critère de validation croisée généralisée (5.8). Il est ensuite possible de calculer la matrice de lissage de l'estimateur de f puis le nombre de degrés de liberté associé.

La description de la méthode est plus générale et peut être étendue à n'importe quelle base de fonctions splines. La simplicité des calculs et le fait qu'elle soit basée sur un critère explicite pour l'estimation sont parmi les raisons de la popularité de cette méthode.

Contrairement à la méthode *loess* le paramètre de lissage des splines de lissage n'a pas une interprétation théorique permettant à l'utilisateur de déterminer la valeur souhaitée. On fixe plutôt le nombre de degrés de liberté. La figure 5.3 représentent trois splines de lissage ayant différentes valeurs de degrés de liberté pour estimer la fonction $f(x) = -x^2$ sur [-3 ; 3]. Elle montre que, plus on diminue le nombre de degrés de liberté, plus l'estimateur est lisse (sa variance est faible alors que son biais varie à nouveau dans le sens inverse).

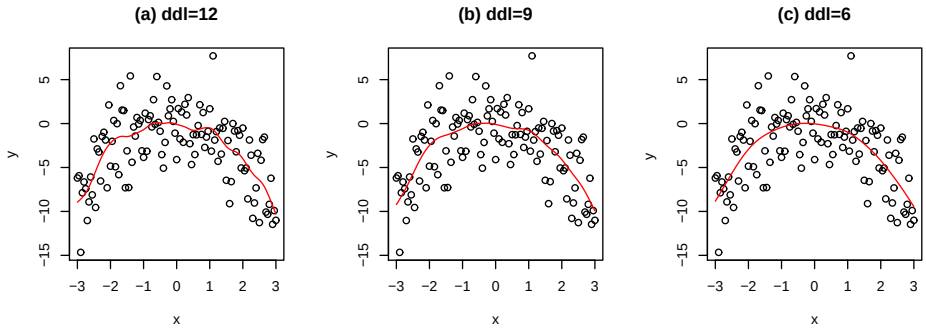

FIG. 5.3 – Illustration des fonctions splines de lissage obtenues en fixant différentes valeurs de degré de liberté sur des données simulées. La fonction à estimer est $f(x) = -x^2$ à laquelle nous avons rajouté un bruit gaussien de moyenne 0 et de variance 3, $x \in [-3, 3]$.

Il existe d'autres méthodes de régression non paramétrique comme les splines de régression, les méthodes à noyaux, la régression par partitionnement, etc. Les splines de régression utilisent souvent des méthodes plutôt intuitives ou d'essais et erreurs pour déterminer l'ensemble des noeuds et par conséquent l'estimateur alors que les splines de lissage déterminent la valeur de ce dernier en minimisant un critère bien précis.

Nous nous limitons à la présentation de la méthode *loess* et les splines de lissage.
L'estimateur donné en (5.23) n'est pas sans biais. Ainsi, il n'est pas possible de faire des tests ou d'établir des intervalles de confiance basés sur cet estimateur. Pour remédier à ces problèmes, une méthode basée sur l'inférence bayésienne a été développée par Wood [42]. Elle sera présentée pour le cas multivarié.

5.3 Régression non paramétrique multivariée

Lorsqu'on s'intéresse aux effets d'un groupe de variables explicatives X_1, X_2, ..., X_p sur une variable réponse Y, on applique une régression multivariée. En général, lorsqu'on dispose d'un échantillon $(y_i, x_{1i}, x_{2i}, ..., x_{pi})$, $i = 1, \ldots, n$, le modèle privilégié, moyennant des transformations sur les variables indépendantes, est le modèle linéaire (5.24).

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i. \tag{5.24}$$

C'est la généralisation du modèle de régression linéaire simple (5.1). Le modèle (5.24) est très simple en termes de calcul et d'interprétation. Il indique que les variables explicatives agissent globalement sur la variable dépendante Y de manière linéaire. Si les corrélations ne sont pas linéaires, on doit transformer les variables. Cette transformation nécessite que l'on ait un fort *a priori* sur la distribution des données.

Pour résoudre ce problème, on peut utiliser les modèles non paramétriques multidimensionnels. Ces modèles sont généralement donnés par (5.25).

$$y_i = f(x_{1i}, x_{2i}, ..., x_{pi}) + \varepsilon_i, i = 1, \ldots, n. \tag{5.25}$$

Le modèle (5.25) ne suppose pas de relation particulière entre la variable réponse et les variables explicatives. Le but de la régression non paramétrique est d'estimer la fonction f.

Plusieurs méthodes de régression non paramétrique multivariée ont été développées ces dernières années. En particulier, toutes les méthodes évoquées dans le cas univarié ont leur version multivariée. Dans cet ensemble, il y a les modèles additifs généralisés. Ces derniers supposent que les variables explicatives sont liées entre elles par une relation additive. Les méthodes d'estimation qui sont appliquées sont les splines de lissage.

5.3.1 Modèles additifs généralisés

Les modèles additifs généralisés ou Generalized Additive Model (GAM) développés par Wood [43], basés sur les travaux de Hastie et Tibshirani [39]

sont une version non paramétrique des modèles linéaires généralisés (GLM) qui utilisent les splines de lissage pour estimer une ou des fonctions de régression non paramétriques. En supposant que les effets des variables explicatives sont indépendants, le modèle (5.25) peut s'écrire sous la forme (5.26). Ce modèle constitue un avantage important du point de vue de l'interprétation des résultats et de la visualisation des fonctions de régression. Les méthodes d'estimation qui sont appliquées sont les splines de lissage mais il existe d'autres méthodes d'estimation. Le choix des splines de lissage est motivé par le fait qu'elles utilisent une méthode bayésienne pour estimer les paramètres ce qui réduit la variance et permettent d'avoir des estimateurs sans biais. Elles permettent donc de faire des tests sur les variables et d'établir des intervalles de confiance.

$$y_i = \alpha_0 + f_1(x_{1i}) + f_2(x_{2i}) + ... + f_p(x_{pi}) + \varepsilon_i, i = 1, \ \ldots, \ n \qquad (5.26)$$

où les résidus ε_i sont des variables aléatoires indépendantes et identiquement distribuées de moyenne nulle et de variance σ^2 et les f_j sont des fonctions inconnues pour lesquelles on impose que les moyennes prises par rapport aux distributions marginales soient toutes nulles.

Dans le cas où on a des variables explicatives catégorielles dans le modèle, on peut réécrire le modèle (5.26) sous la forme (5.27).

$$y_i = \mathbf{x}_i\beta + f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_p(x_{pi}) + \varepsilon_i, i = 1 \ \ldots, \ n \qquad (5.27)$$

où $\mathbf{x}_i$ est le vecteur (ligne) de prédiction associé aux variables catégorielles pour l'observation i et β le vecteur des paramètres associés à ces variables.

5.3.1.1 Estimation des modèles additifs généralisés

Pour estimer les fonctions f_j du modèle (5.27), nous supposons tout d'abord qu'elles sont des fonctions splines. Chaque fonction de lissage f_j peut se décomposer dans une base de fonctions splines qui lui est propre. Pour une base $(b_{jk}(x))_{k=1, \ ..., \ K_j}$ donnée, il existe des coefficients $(\beta^*_{jk})_{k=1, \ ..., \ K_j}$ tels que la décomposition de f_j dans cette base est donnée par :

$$f_j(x_{ji}) = \sum_{k=1}^{K_j} \beta^*_{jk} b_{jk}(x_{ji}) \qquad (5.28)$$

Le modèle (5.27) s'écrit dans ces bases par :

$$y_i = \mathbf{x}_i\beta + \sum_{k=1}^{K_1} \beta^*_{1k} b_{1k}(x_{1i}) + ... + \sum_{k=1}^{K_p} \beta^*_{pk} b_{pk}(x_{pi}) + \varepsilon_i, i = 1, \ \ldots, \ n \qquad (5.29)$$

Dans (5.29), les paramètres inconnus sont le vecteur β et les coefficients $(\beta^*_{jk})_{k=1,\ ...,\ K_j,\ j=1,\ ...,\ p}$. En introduisant les $b_{jk}(x_{ji})$ dans le vecteur ligne $\mathbf{x}_i$ et les β^*_{jk} dans le vecteur colonne β, le modèle (5.29) peut s'écrire sous forme matricielle. Son expression est donnée en (5.30).

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{5.30}$$

C'est un modèle linéaire classique. L'estimateur des paramètres de ce modèle est connu mais la solution souhaitée n'est pas celle qui minimise
$RSS = \|\mathbf{y} - \mathbf{X}\beta\|^2$. Le critère utilisé pour estimer les paramètres avec les splines de lissage est la somme des résidus pénalisée, comme dans le cas univarié. Le lissage est contrôlé par des paramètres qui sont associés aux fonctions f_j à estimer. En s'inspirant du cas univarié, le critère à minimiser devient (5.31).

$$S(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} \lambda_j \int \left(\sum_{k=1}^{K_j} \beta^*_{jk} b''_{jk}(t) \right)^2 dt. \tag{5.31}$$

où λ_j est le paramètre de lissage associé à la fonction f_j. Étant donnée une base de fonctions splines pour chaque fonction f_j, l'intégrale associée à chaque fonction peut s'écrire sous la forme (5.32) comme dans le cas univarié (5.20).

$$\int \left(\sum_{k=1}^{K_j} \beta^*_{jk} b''_{jk}(t) \right)^2 dt = \beta^\top \mathbf{S}_j \beta. \tag{5.32}$$

Les coefficients de la restriction de $\mathbf{S}_j$ par rapport aux paramètres de la fonction f_j (c'est-à-dire le vecteur β^*_j) sont calculés à l'aide de (5.19) et (5.20). Pour avoir une matrice qui a autant de lignes que le vecteur β, on introduit des lignes et des colonnes de 0. Le critère à minimiser est $S(\beta)$ défini en (5.33).

$$S(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \sum_{j=1}^{p} \lambda_j \beta^\top \mathbf{S}_j \beta. \tag{5.33}$$

Le paramètre β minimisant $S(\beta)$ dans (5.33) est donné par :

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X} + \sum_{j=1}^{p} \lambda_j \mathbf{S}_j \right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{5.34}$$

Dans la suite, on pose

$$\mathbf{B} = (\mathbf{X}^\top \mathbf{X} + \textstyle\sum_{j=1}^{p} \lambda_j \mathbf{S}_j)^{-1} \mathbf{X}^\top \text{ et } \hat{\beta} = \mathbf{B}\mathbf{y}$$

Pour espérer réaliser des tests ou construire des intervalles de confiance basés sur l'estimateur $\hat{\beta}$, on peut supposer que les résidus suivent une loi normale multidimensionnelle de moyenne 0 et de variance $\mathbf{I}\sigma^2$. Dans ces conditions la loi de $\hat{\beta}$ est une loi normale multidimensionnelle (5.35).

$$\hat{\beta} \sim \mathcal{N}\left(\mathbb{E}(\hat{\beta}), \mathbf{B}\mathbf{B}^\top\sigma^2\right) \tag{5.35}$$

À partir de (5.34), l'espérance de $\hat{\beta}$ est donnée par :

$$\mathbb{E}(\hat{\beta}) = \left(\mathbf{X}^\top\mathbf{X} + \sum_{j=1}^{p}\lambda_j\mathbf{S}_j\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta \tag{5.36}$$

Dans ces conditions, l'estimateur $\hat{\beta}$ n'est pas sans biais ($\mathbb{E}(\hat{\beta}) \neq \beta$) à moins que le vecteur β soit identiquement nul. Une autre manière de procéder, basée sur l'inférence bayésienne, a été proposée par Wood [43]. Elle fait référence aux travaux de Wahba [44] et Silverman [43] qui supposent que β est une variable aléatoire dont la distribution *a priori* est donnée par (5.37). En fait, l'estimateur obtenu en (5.34) peut être vu comme un estimateur du maximum *a posteriori* avec un *a priori* gaussien (5.37).

$$f_\beta(\beta) \propto \exp\left(-\frac{1}{2}\beta^\top\sum_{j=1}^{p}\frac{1}{\tau_j}\mathbf{S}_j\beta\right) \tag{5.37}$$

où les $\tau_j = \frac{\sigma^2}{\lambda_j}$ sont des paramètres de contrôle de la dispersion de la loi *a priori*.

Conditionnellement à β, (5.30) indique que Y suit une loi normale (5.38).

$$f_{Y|\beta}(\mathbf{y}, \beta) \propto \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)/\sigma^2\right) \tag{5.38}$$

La distribution *a posteriori* de β conditionnellement à Y, obtenue en appliquant le théorème de Bayes et en utilisant les distributions de β et de Y/β est une loi normale. Elle est donnée en (5.39).

$$\beta|Y \sim \mathcal{N}\left(\hat{\beta}, \left(\mathbf{X}^\top\mathbf{X} + \sum_{j=1}^{p}\lambda_j\mathbf{S}_j\right)^{-1}\sigma^2\right) \tag{5.39}$$

À partir de cette expression, des tests et des intervalles de confiance basés sur toute fonction de β peuvent être réalisés. Ils sont fondés sur le fait que $\beta|Y$ suit une loi normale multidimensionnelle de moyenne et de matrice de variance-covariance respectivement définies en (5.40) et (5.41).

$$\mathbb{E}(\beta|Y) = \hat{\beta} = \left(\mathbf{X}^\top \mathbf{X} + \sum_{j=1}^{p} \lambda_j \mathbf{S}_j \right)^{-1} \mathbf{X}^\top \mathbf{y} \tag{5.40}$$

$$\mathbf{V}_\beta = \left(\mathbf{X}^\top \mathbf{X} + \sum_{j=1}^{p} \lambda_j \mathbf{S}_j \right)^{-1} \sigma^2. \tag{5.41}$$

Un des atouts des méthodes GAM vient du fait qu'elles permettent d'ajuster une multitude de modèles différents, spécifiques à chaque situation. Ainsi, pour chaque variable que l'on désire inclure dans le modèle, on peut choisir un ajustement paramétrique ou non paramétrique. Les termes paramétriques peuvent prendre n'importe laquelle des formes utilisées dans les modèles linéaires standards, ce qui implique que l'on peut ajuster des droites pour les covariables ou inclure des variables catégorielles. Le principal avantage de ces méthodes par rapport aux autres méthodes de régression multivariée est qu'elles permettent d'ajuster des modèles non paramétriques simples et faciles à interpréter. Pour chaque variable explicative, on peut choisir le type d'estimation univariée que l'on désire. Toutefois, les logiciels permettant d'ajuster des modèles GAM se limitent habituellement à un petit nombre de fonctions de lissage disponibles. Les fonctions loess et les splines de lissage sont celles qui sont le plus souvent privilégiées. Typiquement, la méthode GAM est utilisée pour ajuster des modèles additifs. Toutefois, il est possible de modifier quelque peu le modèle de départ pour permettre de tenir compte de certaines interactions entre deux variables explicatives. Pour ce faire, on peut par exemple introduire un terme de la forme $f(\mathbf{x}_j,\ \mathbf{x}_l)$ dans le modèle (5.26) et utiliser une fonction de lissage bivariée [39, 42].

5.3.1.2 Calcul du nombre de degrés de liberté

Le but de la pénalisation de l'ajustement est de réduire les variations des degrés de liberté. En l'absence des termes de pénalité dans (5.33), la solution de (5.33) est donnée par :

$$\tilde{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{5.42}$$

Avec la pénalisation, la solution $\hat{\beta}$ de (5.33) sachant Y peut s'écrire de la manière suivante :

$$\begin{aligned} \hat{\beta} &= \left(\mathbf{X}^\top \mathbf{X} + \sum_{j=1}^{p} \lambda_j \mathbf{S}_j \right)^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \left(\mathbf{X}^\top \mathbf{X} + \sum_{j=1}^{p} \lambda_j \mathbf{S}_j \right)^{-1} (\mathbf{X}^\top \mathbf{X})(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

$$= \mathbf{F}\tilde{\beta}$$

avec

$$\mathbf{F} = \left(\mathbf{X}^\top \mathbf{X} + \sum_{j=1}^{p} \lambda_j \mathbf{S}_j \right)^{-1} \mathbf{X}^\top \mathbf{X}$$

La matrice $\mathbf{F}$ permet d'exprimer les paramètres du modèle pénalisé en fonction de ceux du modèle non pénalisé. Ainsi $F_{ii} = \partial\hat{\beta}_i / \partial\tilde{\beta}$ explique comment l'estimateur $\hat{\beta}_i$ du modèle pénalisé peut changer suite à un léger changement des paramètres du modèle non pénalisé. C'est pourquoi F_{ii} mesure les degrés de liberté du $i^{\text{ème}}$ paramètre du modèle pénalisé. Pour le modèle non pénalisé, chaque paramètre a un degré de liberté, mais les termes de pénalisation réduisent la variation des degrés de liberté d'un facteur F_{ii}. La somme des F_{ii} est le nombre de degré de liberté du modèle pénalisé.

En faisant allusion au modèle de régression linéaire, on estime σ^2 par la somme des résidus divisée par le nombre de degré de liberté des résidus (5.43).

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n - \mathrm{tr}(\mathbf{F})}. \tag{5.43}$$

Les paramètres composant le vecteur $\hat{\beta}$ dépendent des λ_j comme $\widehat{\sigma}^2$. Les estimations sont conditionnelles aux λ_j. Pour estimer ces derniers, nous utilisons le critère de validation croisée généralisée. Ce critère minimise les erreurs de prédiction. La fonction à minimiser est donnée en (5.44).

$$V_{GCV}(\lambda) = \frac{n\|Y - X\hat{\beta}\|^2}{[n - \mathrm{tr}(\mathbf{F})]^2}. \tag{5.44}$$

avec $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$.

5.3.1.3 Tests de sous modèles

Un des objectifs d'une étude statistique est de savoir si une ou plusieurs variables explicatives ont des effets significatifs sur la variable réponse. Une autre manière de le dire est de savoir si les informations portées par les variables sont significatives ou non. Pour cela, on réalise des tests d'hypothèses. Supposons qu'on veuille tester la significativité de la variable X_k ($k = 1, \ldots, p$) dans le modèle (5.27). Dans le cas d'un modèle linéaire multidimensionnel (5.24), on peut se restreindre à tester la significativité du paramètre β_k ($\beta_k = 0$ contre $\beta_k \neq 0$) qui est le coefficient de linéarité associé à la variable X_k. En faisant référence à ce cas paramétrique, il est possible de tester la significativité de la fonction f_k pour $k \in \{1, \ldots, p\}$. Ce test est modélisé par (5.45).

$$\begin{cases} H_0 : y_i = \mathbf{x}_i\beta + \sum_{j\neq k} f_j(x_{ji}) + \widetilde{\varepsilon_i} \\ \text{contre} \\ H_1 : y_i = \mathbf{x}_i\beta + \sum_{j=1}^{p} f_j(x_{ji}) + \varepsilon_i \end{cases} \tag{5.45}$$

On note $\beta_j^* = (\beta_{j1}^*, \beta_{j2}^*, \ldots, \beta_{jK_j}^*)^\top$, le vecteur des coefficients de la décomposition de f_j dans la base de fonctions splines (5.29). Pour réaliser le test (5.45), on se sert du vecteur $\hat{\beta}_j^*$ et on teste $\hat{\beta}_j^* = 0$ contre $\hat{\beta}_j^* \neq 0$ (ici 0 désigne le vecteur nul de même longueur que β_j^*). On extrait la matrice de variance-covariance de $\hat{\beta}_j^*$ notée $\mathbf{V}_{\hat{\beta}_j^*}$ dans $\mathbf{V}_\beta$ (5.41). Si $\mathbf{V}_{\hat{\beta}_j^*}$ est de rang plein, alors sous l'hypothèse nulle,

$$\hat{\beta}_j^{*\top}\mathbf{V}_{\hat{\beta}_j^*}^{-1}\hat{\beta}_j^* \sim \chi_d^2$$

ou d=dim(β_j^*). Dans le cas contraire, on note r=rang($\mathbf{V}_{\hat{\beta}_j^*}$) et $\mathbf{V}_{\hat{\beta}_j^*}^{r-}$, la matrice pseudoinverse de $\mathbf{V}_{\hat{\beta}_j^*}$, sous l'hypothèse nulle,

$$\hat{\beta}_j^{*\top}\mathbf{V}_{\hat{\beta}_j^*}^{r-}\hat{\beta}_j^* \sim \chi_r^2$$

La matrice $\mathbf{V}_{\hat{\beta}_j^*}$ dépend de σ^2, estimée dans (5.43). La statistique de test est donnée par (5.46).

$$\hat{\beta}_j^{*\top}\hat{\mathbf{V}}_{\hat{\beta}_j^*}^{r-}\hat{\beta}_j^*/r \sim F_{r,\gamma}. \tag{5.46}$$

où $\gamma = n - \text{tr}(\mathbf{F})$.

Pour tester la significativité d'un paramètre β_j sachant Y, modélisant une modalité d'une variable catégorielle ($\beta_j = 0$ contre $\beta_j \neq 0$), on utilise le test classique de Student. Il est basé sur la statistique T_j définie par :

$$T_j = \frac{\beta_j - \hat{\beta}_j}{\hat{\mathbf{V}}_{\beta_{j,j}}}. \tag{5.47}$$

Sous l'hypothèse nulle, T_j suit une loi de Student à γ degrés de liberté ($\mathcal{T}(\gamma)$). Le test rejette l'hypothèse nulle si $|\hat{T}_j| = \frac{|\hat{\beta}_j|}{\hat{\mathbf{V}}_{\hat{\beta}_{jj}}}$ est supérieur au quantile à $1-\alpha/2$ d'une loi de Student à γ degrés de liberté ou si la p-value observée est inférieure à α.

5.4 Caractérisation des MA et MG

Pour identifier les facteurs liés aux expositions moyennes (MA et MG), nous appliquons les modèles additifs généralisés définis en (5.27). Ces modèles

sont assujettis aux dimensions des bases de fonctions splines pour lesquelles les différentes fonctions sont explicitées (5.28). Dans un contexte de lissage par splines, Kim et Gu (2004) ont montré que les K_j (5.28) devraient être à l'échelle de $n^{2/9}$, où n est le nombre de données. Wood (2006) suggère également que la dimension de la base devrait dépendre du nombre de covariables et la taille de l'échantillon [42]. Il est vrai que le choix des dimensions des bases (les K_j) fait partie intégrante de la spécification du modèle mais la taille exacte de la base n'est pas généralement critique car c'est le paramètre de lissage qui contrôle l'ajustement du modèle. Cet ajustement est donc assez peu sensible à la dimension de la base, pour autant qu'elle n'est pas définie de manière restrictivement faible. Dans le cas ici présent, nous fixons les K_j égaux à 7 et nous estimons les λ_j en minimisant (5.44). La méthode utilisée pour estimer les λ_j est celle de Newton-Raphson.

Pour pouvoir identifier l'effet de chaque variable, nous les introduisons de manière séparée dans le modèle. Cela laisse supposer que les effets des variables continues sur la variable réponse sont indépendantes d'une variable à l'autre. Les variables explicatives considérées sont données dans le tableau 5.1. Elles sont obtenues à partir du questionnaire et de l'emploi du temps.

Variables continues
Densité de population du département
Age
Temps passé dans les transports ferroviaires
Temps passé dans les transports non électriques (voiture, bus, etc.)
Temps passé dans les centres commerciaux
Temps passé sur l'ordinateur
Temps passé devant la télévision
Temps passé à l'école
Temps de sommeil
Variables catégorielles (catégories)
Radio-réveil à moins de 50 cm (Oui/ Non)
Habitation (Appartement/ Pavillon)
Population de la ville de résidence ($>/ \leq$ 2 000 habitants)
Chauffage (Électrique/ Autre)
Type de chauffage (Individuel/ Collectif)
Chauffage d'eau (Électrique/ Autre)
Type de chauffage d'eau (Individuel/ Collectif)
Ligne aérienne à HT à proximité du domicile (Oui/ Non)
Ligne souterraine à HT à proximité du domicile (Oui/ Non)
Réseaux ferrés électrifiés à proximité du domicile (Oui/ Non)
Sexe (Masculin/ Féminin)

TAB. 5.1 – Nom des variables considérées.

Pour prendre en compte les ordres de grandeur et pour des hypothèse de normalité, les moyennes et la densité de population du département sont prises à l'échelle log à base 10. Cette transformation des moyennes permet aussi d'avoir des distributions gaussiennes. Pour les variables catégorielles, nous devons imposer des contraintes d'identifiabilité pour pouvoir estimabler le modèle. Pour cela, nous imposons la nullité d'un paramètre relatif à une modalité pour chaque variable catégorielle (la modalité dont le paramètre est estimé est entre parenthèse dans la colonne "Variable" du tableau du modèle retenu (exemple tableau 5.2).

Lorsque la relation entre l'exposition moyenne (MA ou MG) et une variable est linéaire, nous estimons le coefficient de linéarité et nous l'affichons dans la colonne "Estimation" du tableau du modèle retenu. Dans le cas contraire, la relation est modélisée par une fonction de plus de deux paramètres. Dans ce cas, nous n'affichons que le résultat du test de ces paramètres ou plus précisément de la variable en question. La case estimation reste en blanc pour cette variable. Les tests de significativité sont réalisés en utilisant (5.46) ou (5.47) selon que la variable est continue ou catégorielle. Pour cela, nous

supposons que les résidus du modèle sont indépendants et identiquement distribués selon une loi normale. Seuls les résultats du modèle retenu sont représentés.

5.4.1 Exposition sur 24 heures

5.4.1.1 Les enfants

Les variables retenues sont données dans les tableaux 5.2 et 5.3. Ces tableaux montrent que les moyennes arithmétiques et géométriques sont plus élevées chez les enfants qui ont posé l'EMDEX à proximité du radioréveil, qui habitent dans un appartement et qui résident dans une ville de plus de 2 000 habitants. Ces moyennes sont aussi plus élevées chez les enfants qui ont leurs foyers à côté des lignes aériennes à haute tension ou des réseaux ferrés électrifiés. Par contre, elles diminuent avec le temps passé à l'école. D'autres variables apparaissent aussi comme facteurs d'exposition comme la densité de population pour les deux moyennes, l'âge pour les MA (les MA croissent avec l'âge). Pour les MG, on a identifié le temps passé sur ordinateur (les MG croissent avec le temps passé sur ordinateur), le temps passé dans les transports ferroviaires et dans les transports non électriques. Les taux de variance expliquée sont de $17,2\%$ pour les MA et $27,2\%$ pour les MG : les modèles ne sont pas prédictifs.

Variable	Estimation	P-value
Age	0,02	0,001
Temps passé à l'école	-0,02	0,005
Radio-réveil (Oui)	0,42	$< 0,001$
Habitation (Appartement)	0,12	0,017
Population ($>$ 2 000 habitants)	0,10	0,017
Lignes aériennes à HT (Oui)	0,34	0,004
Réseaux ferrés électrifiés (Oui)	0,21	0,001
Densité de population du département		$< 0,001$

TAB. 5.2 – Variables explicatives retenues pour les MA des enfants.

Variable	Estimation	P-value
Temps passé sur l'ordinateur	0,18	$< 0{,}001$
Temps passé à l'école	-0,03	0,036
Radio-réveil (Oui)	0,62	$< 0{,}001$
Habitation (Appartement)	0,57	$< 0{,}001$
Population ($> 2\,000$ habitants)	0,41	$< 0{,}001$
Ligne aériennes à HT (Oui)	0,98	0,002
Réseaux ferrés électrifiés (Oui)	0,73	$< 0{,}001$
Temps passé dans les transports ferroviaire		0,039
Temps passé dans les transports non électriques		0,006
Densité du département		$< 0{,}001$

TAB. 5.3 – Variables explicatives retenues pour les MG des enfants.

5.4.1.2 Les adultes

Les moyennes croissent linéairement avec le temps passé sur ordinateur et dans les centres commerciaux (tableaux 5.4 et 5.5). Elles sont plus élevées pour les adultes qui ont leurs foyers à proximité des lignes aériennes à HT et ou qui ont posé l'EMDEX à côté du radio-réveil. D'autres variables peuvent être considérées comme facteurs d'exposition selon le type de moyenne comme le temps passé dans les transports ferroviaires pour les MA ou le fait d'avoir son foyer de résidence dans une ville de plus de 2 000 habitants ou d'habiter dans un appartement pour les MG. Comme dans le cas des enfants, les modèles retenus ne sont pas prédictifs. Les taux de la variance expliquée par ces modèles sont respectivement de $16,7\%$ pour les MA et $24,5\%$ pour les MG.

Variable	Estimation	P-value
Temps passé dans les transports ferroviaires	0,33	$< 0{,}001$
Temps passé dans les centres commerciaux	0,13	0,001
Temps passé sur l'ordinateur	0,23	0,013
Radio-réveil (Oui)	0,84	$< 0{,}001$
Lignes aériennes à HT (Oui)	0,89	0,001
Réseaux ferrés électrifiés (Oui)	0,31	0,015
Densité de population du département		$< 0{,}001$

TAB. 5.4 – Variables explicatives retenues pour les MA des adultes.

Variable	Estimation	P-value
Temps passé dans les centres commerciaux	0,15	0,001
Temps passé sur l'ordinateur	0,04	0,001
Radio-réveil (Oui)	0,62	$< 0,001$
Population ($> 2\,000$ habitants)	0,42	$< 0,001$
Habitation (Appartement)	0,59	$< 0,001$
Lignes aériennes à HT (Oui)	1,55	$< 0,001$
Réseaux ferrés électrifiés (Oui)	0,58	$< 0,001$
Densité de population du département		$< 0,001$

TAB. 5.5 – Variables explicatives retenues pour les MG des adultes.

5.4.2 Exposition hors période de sommeil

Pour les enfants, en supprimant les CM enregistrés pendant le sommeil, nous éliminons principalement les effets des radio-réveils et de nouveaux facteurs apparaissent de manière significative. Ces derniers sont le temps passé sur ordinateur pour les deux moyennes, le temps passé dans les transports ferroviaires pour les MA et celui passé dans les centres commerciaux pour les MG (tableaux 5.6 et 5.7). Toutefois les taux de variance expliquée sont du même ordre ($17,7\%$ pour le modèle des moyennes arithmétiques et $29,8\%$ pour celui des moyennes géométriques).

Pour les adultes, les moyennes arithmétiques décroissent linéairement avec l'âge alors que les moyennes géométriques augmentent avec le temps passé dans les transports ferroviaires (tableaux 5.8 et 5.9). Les taux de variance expliquée sont aussi très faibles. Ils sont de $12,5\%$ pour le modèle associé aux MA et $22,2\%$ pour celui des MG.

Variable	Estimation	P-value
Habitation (Appartement)	0,21	0,030
Population ($> 2\,000$ habitants)	0,20	0,006
Lignes aériennes à HT (Oui)	0,73	0,001
Réseaux ferrés électrifiés (Oui)	0,38	0,001
Age		$< 0,001$
Temps passé dans les transports ferroviaires		0,001
Temps passé sur l'ordinateur	0,06	0,027
Temps passé à l'école		0,001
Densité de population du département		$< 0,001$

TAB. 5.6 – Variables explicatives retenues pour les MA des enfants.

Variable	Estimation	P-values
Population (> 2 000 habitants	0,32	< 0,001
Habitation (Appartement)	0,34	0,001
Lignes aériennes à HT (Oui)	0,67	0,002
Réseaux ferrés électrifiés (Oui)	0,49	< 0,001
Temps passé dans les transports ferroviaires		0,001
Temps passé dans les transports non électriques		0,003
Temps passé dans les centres commerciaux		0,014
Temps passé devant la télévision	0,05	0,033
Temps passé sur l'ordinateur	0,13	< 0,001
Temps passé à l'école	-0,04	< 0,001
Densité de population du département		< 0,001

TAB. 5.7 – Variables explicatives retenues pour les MG des enfants.

Variable	Estimation	P-value
Lignes aériennes à HT (Oui)	0,26	0,002
Réseaux ferrés électrifiés (Oui)	0,12	0,015
Age	-0,01	0,010
Temps passé dans les transports ferroviaires		< 0,001
Temps passé dans les centres commerciaux	0,05	< 0,001
Temps passé sur l'ordinateur	0,01	0,042
Densité de population du département		< 0,001

TAB. 5.8 – Variables explicatives retenues pour les MA des adultes.

Variable	Estimation	P-value
Population (> 2 000 habitants)	0,10	0,001
Habitation (Appartement)	0,15	0,002
Lignes aériennes à HT (Oui)	0,45	< 0,001
Réseaux ferrés électrifiés (Oui)	0,19	< 0,001
Temps passé dans les transports ferroviaires	0,08	0,008
Temps passé dans les centres commerciaux	0,06	< 0,001
Temps passé sur l'ordinateur		< 0,001
Densité de population du département		< 0,001

TAB. 5.9 – Variables explicatives retenues pour les MG des adultes.

5.5 Conclusion

Dans ce chapitre, nous avons identifié des facteurs favorisant l'exposition en termes de moyennes arithmétiques et géométriques. Ils sont associés à différentes variables continues comme le temps passé dans les transports ferroviaires, les centres commerciaux, le temps passé sur ordinateur ou encore la densité du département de résidence. Pour les variables catégorielles, nous avons trouvé que l'exposition est plus élevée pour les personnes qui résident dans les grandes villes, qui habitent dans des appartements et/ou qui ont leurs foyers de résidence à côté des lignes aériennes à haute tension ou des réseaux ferrés électrifiés. Les radio-réveils sont peut être la source de champ la plus importante car elle peut générer des CM de plus forte intensité (jusqu'à plusieurs dizaines de micro Teslas au contact) mais nous avons vu au chapitre 4 que nous ne pouvons pas dire s'ils influencent l'exposition ou non.

Nous retrouvons les mêmes facteurs favorisant les moyennes les plus élevées identifiés dans le chapitre 4. Par contre les câbles souterrains n'apparaissent pas comme étant des facteurs d'exposition.
Les modèles utilisés pour caractériser les moyennes arithmétiques et géométriques ont permis de montrer, à l'aide de tests statistiques, que les structures de dépendances entre ces moyennes et certaines variables ne sont pas linéaires. Malheureusement les modèles non linéaires testés ne permettent pas d'obtenir de bons taux de variance expliquée. Il semblerait donc que les facteurs disponibles dans la base de données ne suffisent pas à expliquer à eux seuls l'exposition moyenne d'un individu. Ces taux sont légèrement plus élevés pour les modèles relatifs aux moyennes géométriques qui sont moins sensibles aux valeurs extrêmes des CM par rapport à ceux des moyennes arithmétiques.

Chapitre 6

Recherche de classes d'exposition

6.1 Introduction

Pour caractériser les expositions, on s'est intéressé au chapitre précédant aux moyennes arithmétique et géométrique. Or les séries de champs magnétiques ne sont pas stationnaires, ces deux moyennes ne permettent pas à elles seules de bien résumer les séries de champs magnétiques. C'est pour cette raison que nous nous intéressons à des classes d'expositions qui seront établies à l'aide de plusieurs descripteurs de chaque série.
Dans ce chapitre, on se donne comme but de regrouper les individus ayant les séries de CM les plus proches. La proximité entre les séries de CM sera traduite par une distance. On dit aussi qu'on veut faire de la classification non supervisée. L'objectif d'une classification non supervisée est la recherche d'une répartition des individus en classes ou catégories. Ceci est fait en optimisant un critère visant à regrouper les individus dans des classes les plus homogènes possible et les plus distinctes possibles entre elles. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement, pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage. On est ici dans une situation d'apprentissage non supervisé (clustering en anglais). Il existe plusieurs méthodes de classification non supervisée. Nous nous sommes intéressés à la classification ascendante hiérarchique (CAH) car nous n'avons pas d'*a priori* ou de connaissance sur le nombre de classes à retenir et en plus nous n'avons pas beaucoup de données. Il s'agit de regrouper itérativement les individus, en supposant au départ que chaque individu forme un groupe et en construisant progressivement un arbre, ou dendrogramme, regroupant finalement tous les individus en une seule classe (le nombre de classes est déterminé *a posteriori*, à la vue du dendrogramme). Ceci suppose de savoir calculer, à chaque étape ou regroupement, la distance entre un individu et un groupe ainsi que celle entre deux groupes. Ceci nécessite donc, pour l'utilisateur de cette méthode de définir une distance entre deux groupes connaissant celles de tous les couples d'individus de ces deux groupes. Il existe plusieurs distances, appelées aussi saut ou encore linkage en anglais.

6.2 Mesure d'éloignement

Pour réaliser une CAH, on doit définir ou choisir une distance pour pouvoir caractériser le rapprochement ou l'éloignement entre deux individus ou d'un individu à une classe d'individus.

Notons $\mho = \{1, 2, \ldots, n\}$, l'ensemble des individus. Nous allons définir sur $\mho \times \mho$ les mesures d'éloignement entre deux individus ou deux classes d'individus utilisées dans cette étude.

6.2.1 Distance euclidienne

Une distance d définie dans $\mho \times \mho$ est une application de $\mho \times \mho$ dans $\mathbb{R}_+$ vérifiant (6.2), (6.3) et (6.3).

$$\begin{aligned} d(i,j) &= d(j,i),\ \forall (i,j) \in \mho \times \mho. && (6.1)\\ d(i,j) &= 0 \Leftrightarrow i = j,\ \forall (i,j) \in \mho \times \mho. && (6.2)\\ d(i,j) &\leq d(i,k) + d(k,j),\ \forall (i,j,k) \in \mho^3. && (6.3) \end{aligned}$$

Lorsque $\mho$ est un espace vectoriel muni d'un produit scalaire $(< \ . \ >)$, donc d'une norme, la distance définie à partir de cette norme appelée aussi distance euclidienne est définie par (6.4).

$$d(i,j) =< i-j, i-j >^{1/2} = \|i-j\|,\ \forall (i,j) \in \mho \times \mho. \qquad (6.4)$$

6.2.2 Distance entre deux classes

Soient A et B, deux classes ou éléments d'une répartition donnée, ω_A et ω_B leurs pondérations (nombres d'individus de chaque classe).
On définit les coordonnées des barycentres G_A et G_B par les moyennes des coordonnées des individus de chaque classe.
À partir de ces barycentres, on va étendre la distance entre deux points à la distance entre ces deux classes par celle des barycentres ou par le saut de Ward donné par la formule (6.5).

$$d(A,B) = \frac{\omega_A \omega_B}{\omega_A + \omega_B} d(G_A, G_B). \qquad (6.5)$$

où $d(G_A, G_B)$ désigne la distance entre les barycentres G_A et G_B.

Le saut de Ward joue un rôle particulier et est la stratégie la plus courante. En effet ce critère induit, à chaque étape de regroupement, une minimisation de la variance interclasses.

6.3 Algorithme de la CAH

Il se résume en deux points :
- Initialisation : Les classes initiales sont les singletons (chaque individu constitue un groupe). Calculer la matrice de leurs distances deux à deux.
- Itération : Itérer les deux étapes suivantes jusqu'à l'agrégation en une seule classe :
  1. Regrouper les deux classes les plus proches au sens de la distance entre les classes choisies.

2. Mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes.

Pour comparer les MA et les MG des classes retenues, nous allons réaliser des tests d'égalité des paramètres de localisation. Pour cela, nous appliquons le test de Moses pour comparer les paramètres de dispersion en calculant un intervalle de confiance à 95% de la p-value observée sur 100 réalisations. Si le seuil α appartient à cet intervalle de confiance, on applique le test de Fligner-Policello sinon celui de Wilcoxon-Mann-Whitney est utilisé pour comparer les paramètres de localisation.

6.4 Caractérisation des classes des plus exposées

Une fois la classification réalisée, on aimerait décrire en particulier les classes regroupant les individus les plus exposés que nous appellerons dans la suite « classes des plus exposés » c'est-à-dire connaître les variables qui peuvent conduire un individu à appartenir à ces classes. Pour cela, nous réalisons une régression logistique.

L'analyse de régression est une technique statistique permettant d'établir une relation entre une variable dépendante et des variables explicatives, afin d'étudier des associations et faire des prévisions. Lorsque la variable dépendante n'est pas quantitative mais qualitative ou catégorielle, le modèle de régression linéaire n'est plus approprié. On utilise alors le modèle de régression logistique. Il s'agit d'un modèle comparable au modèle de régression linéaire, sauf que la variable à expliquer est catégorielle c'est-à-dire que cette variable ne peut prendre que des attributs. On peut, par exemple, s'intéresser à quantifier la relation entre le risque de décès et la quantité de cigarettes fumées par jour, tout en ajustant pour l'âge, le sexe et éventuellement d'autres facteurs de risque. Dans le cas d'un modèle de régression linéaire, la variable à expliquer est en revanche quantitative. Dans ce cas, l'hypothèse de normalité de la distribution des erreurs de cette variable ou d'une transformation de cette variable peut être acceptable, tandis que lorsqu'elle est qualitative elle n'admet pas de valeur numérique naturelle (puisqu'elle ne peut prendre que des attributs) et le modèle normal n'est pas approprié. Une variable aléatoire qualitative est décrite par les probabilités des différents attributs qu'elle peut prendre et pour évaluer l'influence de différents facteurs sur cette variable, il est d'usage de modéliser les probabilités des différents attributs.

6.4.1 Le modèle de régression logistique

Cette section décrit la modélisation d'une variable qualitative Z à 2 modalités 1 ou 0, succès ou échec, présence ou absence de maladie, bon ou mauvais client, etc. Le modèle est adapté à cette situation en cherchant à expliquer les probabilités $\pi = \mathbb{P}(Z = 1)$ (ou $1-\pi = \mathbb{P}(Z = 0)$) ou plutôt une transformation de celles-ci. Le modèle est calibré à l'aide de l'observation conjointe des variables explicatives. L'idée est donc de faire intervenir une fonction réelle monotone g, appelée fonction de lien, opérant de [0, 1] dans $\mathbb{R}$ et donc de chercher un modèle linéaire de la forme (6.6).

$$g(\pi_i) = \mathbf{x}_i^\top \theta. \tag{6.6}$$

où $\mathbf{x}_i$ est le vecteur contenant les variables explicatives de l'individu i et θ le vecteur des paramètres.

Il existe de nombreuses fonctions dont le graphe présente une forme sigmoïdale et qui peuvent remplir ce rôle. Trois d'entre elles sont en pratique utilisées dans les logiciels :

- Le lien *probit* : g est la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.
- Le lien log-log : g est dans ce cas définie par $g(\pi) = \log(1-\log(1-\pi))$
- Le lien *logit* : la fonction g est définie par :

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \tag{6.7}$$

Une des raisons qui font préférer le lien *logit* est que le rapport $\frac{\pi}{1-\pi}$ qui exprime une "cote" est un odds ratio (OR) et la régression logistique s'interprète donc comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio" (l'influence d'un facteur qualitatif sur le risque en question). Dans toute la suite, g désigne la fonction définie en (6.7).

Pour chaque individu i, on réalise une observation (cas de données individuelles) de la variable Z notée z_i. On suppose que toutes les observations sont indépendantes conditionnellement aux covariables ou variables explicatives et on note π_i la probabilité de succès pour l'individu i. Alors, la variable Z_i d'espérance $\mathbb{E}(Z_i) = \pi_i$ suit une loi de Bernoulli $\mathfrak{B}(\pi_i)$. Sa fonction de densité est donnée en (6.8) et (6.9).

$$\mathbb{P}(Z = z_i) = \pi_i^{z_i}(1-\pi_i)^{1-z_i},\ i = 1,\ \ldots,\ n. \tag{6.8}$$

$$g(\pi_i) = \mathbf{x}_i^\top \theta,\ i = 1,\ \ldots,\ n. \tag{6.9}$$

Les probabilités π_i sont données par la formule (6.10).

$$\pi_i = \frac{\exp(\mathbf{x}_i^\top \theta)}{1 + \exp(\mathbf{x}_i^\top \theta)},\ i = 1,\ \ldots,\ n. \tag{6.10}$$

6.4.2 Estimation des paramètres du modèle

Pour estimer les paramètres θ du modèle formé par les équations (6.8) et (6.9), on utilise la méthode de maximum de vraisemblance. Elle consiste à écrire la vraisemblance du modèle et à estimer les paramètres qui maximisent la log-vraisemblance.
De 6.8, on peut écrire :

$$\mathbb{P}(Z = z_i) = \exp\left(z_i \log\left(\tfrac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right)$$

Et de (6.9), on a .

$$\mathbb{P}\left(Z = z_i\right) = \exp\left(z_i \mathbf{x}_i^\top \theta - \log\left(1 + \exp(\mathbf{x}_i^\top \theta)\right)\right) \tag{6.11}$$

En supposant que θ appartienne à un espace de dimension finie p, on peut écrire la vraisemblance $l(\theta)$ du modèle (6.12). À partir de (6.12), on calcule la log-vraisemblance $\mathcal{L}(\theta)$ (6.13).

$$l(\theta) = \prod_{i=1}^{n} \exp\left(z_i \mathbf{x}_i^\top \theta - \log\left(1 + \exp(\mathbf{x}_i^\top \theta)\right)\right) \tag{6.12}$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \left(z_i \mathbf{x}_i^\top \theta - \log\left(1 + \exp(\mathbf{x}_i^\top \theta)\right)\right) \tag{6.13}$$

L'estimateur $\hat{\theta}_n$ du maximum de vraisemblance de θ est solution du système d'équations (6.14). Sous des hypothèses de régularité adaptées (θ appartient dans un espace ouvert convexe, la fonction de lien g est 2 fois continûment différentiable, condition pour que la matrice hessienne de $\mathcal{L}$ soit définie positive), $\hat{\theta}_n$ existe et est un estimateur consistant. On note $\mathcal{I}_n(\theta)$ l'information de Fisher du modèle. Elle est définie en (6.15).

$$\frac{\partial \mathcal{L}}{\partial \theta_j}(\hat{\theta}_n) = 0,\ \forall\ j = 1,\ \ldots,\ p \tag{6.14}$$

$$\mathcal{I}_n(\theta) = -\mathbb{E}\left(\frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_k}(\theta)\right),\ 1 \leq j, k \leq p \tag{6.15}$$

Le système (6.14) n'a pas une solution explicite. Les logiciels calculent les estimations en utilisant un algorithme itératif pour la resolution de ces équations non linéaires comme celui de Newton-Raphson. Il existe une variante de cet algorithme appelé algorithme de « Fisher scoring ». Pour cela, on pose $U(\theta) = \frac{\partial \mathcal{L}}{\partial \theta}(\theta)$ le vecteur des scores et $\mathbf{H}(\theta) = \frac{\partial^2 U}{\partial \theta_j \partial \theta_k}(\theta)$ la matrice

hessienne de $\mathcal{L}$ par rapport à θ.
En supposant que $\hat{\theta}_n$ est un estimateur consistant de θ, on a :

$$U(\hat{\theta}_n) \approx U(\theta) + \mathbf{H}(\theta)(\hat{\theta}_n - \theta)$$

Et par définition, $U(\hat{\theta}_n) = 0$ d'où (6.16).

$$\hat{\theta}_n \approx \theta - \mathbf{H}^{-1}(\theta)U(\theta). \tag{6.16}$$

On définit ainsi la méthode itérative de Fisher scoring (6.17)qui converge vers θ et qui fournit un estimateur de la variance de $\hat{\theta}_n$ par (6.18). Pour l'application de cet algorithme, on estime $\mathcal{I}_n(\theta)$ par $-\mathbf{H}(\hat{\theta})$.

$$\hat{\theta}^{(r+1)} = \hat{\theta}^{(r)} + \hat{\mathcal{I}}_n^{-1}(\hat{\theta}^{(r)})U(\hat{\theta}^{(r)}). \tag{6.17}$$

La variance de $\hat{\theta}_n$ est ensuite estimée par :

$$\hat{\mathbb{V}}(\hat{\theta}_n) = \hat{\mathcal{I}}_n^{-1}(\hat{\theta}_n) \tag{6.18}$$

L'algorithme approxime le logarithme de la fonction de vraisemblance, dans un voisinage du paramètre initial par une fonction qui a la forme d'une parabole concave. Il utilise la matrice de l'information de Fisher contenant l'information concernant la courbure de la fonction log-vraisemblance au point d'estimation. Plus grande est la courbure, plus l'information apportée aux paramètres est importante. Les écart-types des estimateurs sont les racines carrées des éléments diagonaux de l'inverse de l'information de Fisher. Plus la courbure de la fonction de log-vraisemblance est importante plus les écart-types sont petits car une grande courbure implique que la log-vraisemblance diminue rapidement quand on s'éloigne de $\hat{\theta}_n$. En conséquence les données observées ont plus de chance d'apparaître si θ prend la valeur de $\hat{\theta}_n$ plutôt qu'une valeur éloignée de $\hat{\theta}_n$.

6.4.3 Test de significativité des paramètres

Pour réaliser les tests de significativité des paramètres, on se sert du fait que $\mathcal{I}_n^{1/2}(\hat{\theta}_n)(\hat{\theta}_n - \theta)$ converge vers une loi normale p-dimensionnelle centrée et réduite. Ainsi, pour tester l'hypothèse $H_0 : \theta_j = 0$ contre $H_0 : \theta_j \neq 0$, on utilise le fait que $\frac{\hat{\theta}_{n,j} - \theta_j}{\sqrt{\hat{\mathbb{V}}(\hat{\theta}_n)_{j,j}}} \rightarrow \mathcal{T}(n-p)$, sous H_0 avec p la dimension de θ.

Le test rejette H_0, au seuil α, si $T_{obs} = \frac{|\hat{\theta}_{n,j}|}{\sqrt{\hat{\mathbb{V}}(\hat{\theta}_n)_{j,j}}}$ est supérieure au quantile à $1 - \alpha/2$ d'une loi de Student ou tout simplement si la p-value est inférieure à α.

6.4.4 Test de sous modèles

Pour réaliser des tests de sous modèles, nous allons utiliser le test de rapport des vraisemblances. Soit M_p un modèle à p paramètres défini comme dans les équations (6.8) et (6.9) et soit $L_{mv}(M_p)$ le maximum de vraisemblance sous le modèle M_p. Soit $M_{p+p'}$ un modèle à $p+p'$ paramètres et M_p le sous modèle de $M_{p+p'}$ obtenu en éliminant p' covariables ou facteurs. On veut tester l'influence des p' covariables sur le modèle. L'apport de ces covariables est mesuré à l'aide de la statistique du rapport des vraisemblances T_{RV} donnée en (6.19).

$$T_{RV} = -2\left(\log(L_{mv}(M_p)) - \log(L_{mv}(M_{p+p'}))\right) \tag{6.19}$$

Sous l'hypothèse du modèle M_p, T_{RV} suit une loi de χ^2 à p' degrés de liberté. Le test rejette le modèle M_p si le rapport des vraisemblances observé $\hat{T}_{RV}$ est supérieur au quantile à $1-\alpha$ d'une loi de $\chi^2(p')$ ou si la p-value est inférieure à α.

6.5 Application de la classification ascendante hiérarchique

Dans cette partie nous allons chercher des classes ou groupes d'exposition. Le but est d'avoir des classes ou groupes d'individus ayant des caractéristiques d'expositions très proches (aucune hypothèse sur les données n'est requise). Pour cela, nous résumons chaque série de CM par un ou des descripteurs traduisant :

- les valeurs moyennes (la moyenne arithmétique, la moyenne géométrique et la médiane)
- les valeurs élevées (le troisième quantile et la valeur maximale)
- le niveau de variation autour de la moyenne (l'écart-type)
- la vitesse de variation (le RCMS ou Rate Change of Metric Standardized (6.20)).

$$RCMS = \frac{1}{\hat{\sigma}}\sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}(x_t - x_{t+1})^2}. \tag{6.20}$$

où $(x_t)_{t=1,2,\ldots,T}$ est la série de CM et $\hat{\sigma}$ son écart-type.

Le RCMS est un indicateur mesurant la stabilité temporelle de la série de CM.

Ces descripteurs sont centrés et réduits. Le choix des descripteurs n'est pas exhaustif mais nous supposons qu'en considérant ces 7 descripteurs, on peut mieux décrire comment est exposé chaque individu. Chaque série de

CM est donc caractérisée par les valeurs des descripteurs qui lui sont associées.

Le résultat de la CAH est un dendrogramme donnant la formation ou l'agrégation des classes. Pour décider du nombre de classes à retenir, nous représentons la décroissance de la variance interclasses en fonction du nombre de classes considéré. La présence d'une rupture importante dans cette décroissance aide au choix du nombre de classes comme dans le cas du choix de dimension en analyse en composantes principales, avec la décroissance des valeurs propres. Ce graphe se lit de droite à gauche et on s'arrête avant le premier saut jugé significatif. L'indice de Ward par exemple, revient à couper le dendrogramme avant une perte, jugée trop importante de la variance interclasses.

Pour identifier les facteurs qui caractérisent la probabilité d'appartenir à la classe des plus exposés, nous appliquons une régression logistique. A partir du modèle contenant l'ensemble des variables, nous réalisons une sélection des variables les plus significatives en réalisant des tests de significativité ou de sous modèles. Seuls les résultats du modèle retenu sont présentés dans la suite.

6.5.1 Classification selon les CM enregistrés sur 24 heures

6.5.1.1 Les enfants

Le dendrogramme et la courbe d'aide à la décision sont donnés dans la figure 6.1. Cette figure montre qu'on commence à perdre de la variance interclasses de manière significative lors de l'agrégation de quatre classes en trois classes. On peut donc se décider de garder quatre classes mais la troisième classe est moins représentée (coupure en noir), elle compte 8 personnes. Nous avons décidé de la fusionner avec celle qui lui est la plus proche au sens du saut de Ward (coupure en rouge). Nous retenons ainsi trois classes d'exposition.

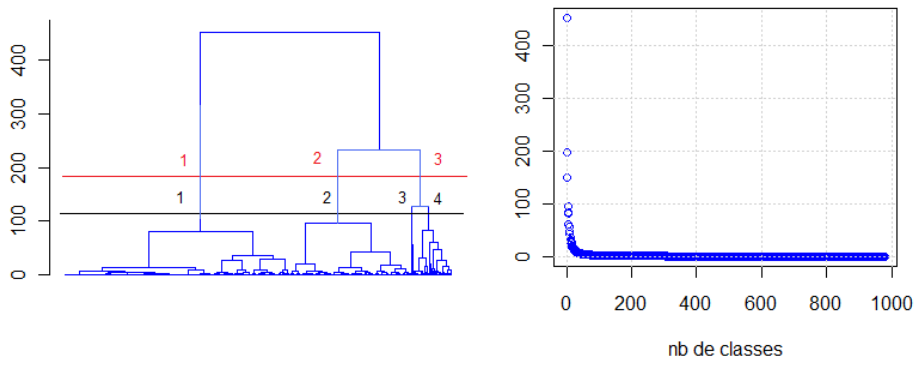

FIG. 6.1 – Dendrogramme de la classification des enfants (figure de gauche) et décroissance de la variance interclasses (figure de droite).

Le tableau 6.1 donne les expositions moyennes des individus de chaque classe en termes de moyennes arithmétique et géométrique. Il montre que la classe 3 est la moins représentée, elle est composée des individus les plus exposés (la moyenne arithmétique observée dans cette classe est 10 fois plus élevée que celle enregistrée dans les deux autres). Le test d'égalité des paramètres de localisation des MA des classes 1 et 2 ne rejette pas cette hypothèse avec une p-value=0,126 (le test réalisé est celui de Wilcoxon-Mann-Whitney car le test de Moses ne rejette pas l'égalité des paramètres de dispersion avec une p-value moyenne de 0,294, IC=[0,256 ; 0,333]). Pour les classes 2 et 3, cette hypothèse est rejetée en faveur d'une exposition plus élevée pour la classe 3 avec une p-value inférieure à 0,001. La figure 6.2 donne la position des fonctions de répartition des MA observées par les individus des trois classes. Elle montre effectivement que celle de la classe 3 est excentrée vers la droite par rapport à celle de la classe 2. Pour les MG, les résultats des tests indiquent que la classe 2 est moins exposée que la première alors que la classe 3 est plus exposée que la première. On peut dire que les classes 1 et 2 ne sont pas très différentes au sens des MA et des MG. La différence apparait avec la classe 3.

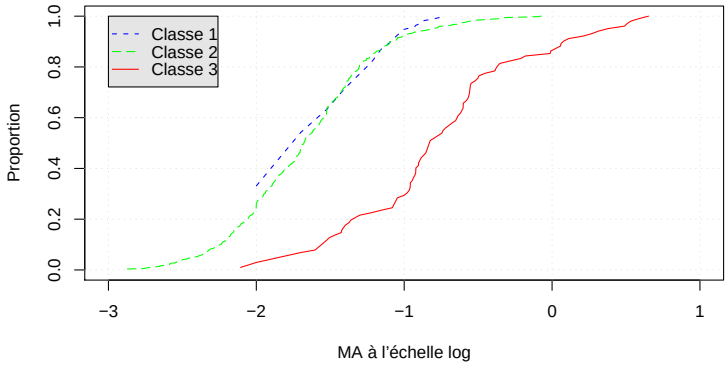

FIG. 6.2 – Fonction de répartition des MA des trois classes.

Classe	Classe 1	Classe 2	Classe 3
Nb de personnes	308 (31,5%)	568 (58,1%)	102 (10,4%)
Moyenne des MA en μT (valeur min ; valeur max)	0,036 (0 ; 0,19)	0,043 (0 ; 0,85)	0,473 (0,01 ; 4,49)
Moyenne des MG en μT (valeur min ; valeur max)	0,020 (0 ; 0,1)	0,008 (0 ; 0,07)	0,082 (0 ; 0,78)

TAB. 6.1 – Expositions moyennes des enfants de chaque classe en considérant les CM sur 24 heures.

La modélisation de la probabilité d'appartenir à la classe 3 montre que cette probabilité diminue avec le temps passé à l'école et dans les transports non électriques. Elle est par contre plus élevée chez les enfants ayant posé l'EMDEX à proximité du radio-réveil ou qui habitent dans des foyers proches des réseaux ferrés électrifiés (Tableau 6.2). Les odds ratio associés à ces modalités sont respectivement de 2,1 (IC=[1,3 ; 3,5]) et 2,3 (IC=[1,2 ; 4,2]). Un enfant ayant posé l'EMDEX à proximité du radio-réveil ou qui habite à côté d'un réseau ferré électrifié a pratiquement 2 fois plus de probabilité d'appartenir à la classe des plus exposés qu'un enfant qui a suivi les instructions en éloignant l'EMDEX du radio-réveil ou qui n'a pas son foyer à côté d'un réseau ferré électrifié. Ces résultats reflètent bien le fait que les enfants sont plus exposés au domicile qu'à l'extérieur. Les réseaux à haute tension n'apparaissent pas dans les variables significatives.

Variable	Estimateur	Écart-type	p-value
Temps passé dans les transports non électriques	-0,42	0,18	0,017
Temps passé à l'école	-0,08	0,03	0,009
Radio-réveil (Oui)	0,76	0,25	0,003
Réseaux ferrés électrifié (Oui)	0,82	0,31	0,007

TAB. 6.2 – Variables retenues comme significative pour la modélisation de la probabilité d'appartenir à la classe des plus exposés pour les enfants en considérant les CM sur 24 heures.

6.5.1.2 Les adultes

Le nombre de classes retenues est 3 (figure 6.4). Le tableau 6.3 donne les expositions moyennes observées dans chaque classe. Il montre que la classe des moins exposés est composée de $74,3\%$ des adultes. Les moyennes arithmétique et géométriques observées dans cette classe sont très faibles (0,052 μT pour les MA et 0,014 μT pour les MG). Les moyennes des MA et des MG observées dans la classe 2 sont respectivement cinq fois et six fois plus élevées que celles enregistrées dans la classe 1. Les moyennes de cette classe composée de $22,9\%$ des adultes sont par contre cinq fois et deux fois moins élevées que celles calculées dans la classe 3 respectivement pour les MA et les MG.

Le test de Moses d'égalité des paramètres de dispersion appliqué sur les MA et les MG des classes 1 et 2 rejette cette hypothèse (les bornes supérieures des intervalles de confiance des p-values sont inférieures à 0,001). Ainsi, nous avons appliqué le test de Wilcoxon-Mann-Whitney pour comparer les paramètres de localisation des MA et des MG. En stipulant que les fonctions de répartition sont identiques dans les deux classes contre des expositions plus élevées dans la classe 2 que dans la classe 1, nous avons trouvé des p-values inférieures à 0,001. Ces tests montrent que les individus de la classe 2 sont plus exposés que ceux de la classe 1.
Nous avons supposé que les paramètres de dispersion des MA et des MG sont aussi différents dans les classes 2 et 3 (la taille de la classe 3 est faible et on ne peut pas s'assurer de la convergence de la statistique de Moses). Nous appliquons le test de Wilcoxon-Mann-Whitney pour comparer les paramètres de localisation des moyennes. Ce test a montré que les MA et les MG de la classe 3 sont globalement plus élevées que celles de la classe 2. Les fonctions de répartition empirique des MA des trois classes sont représentées dans la figure 6.3.

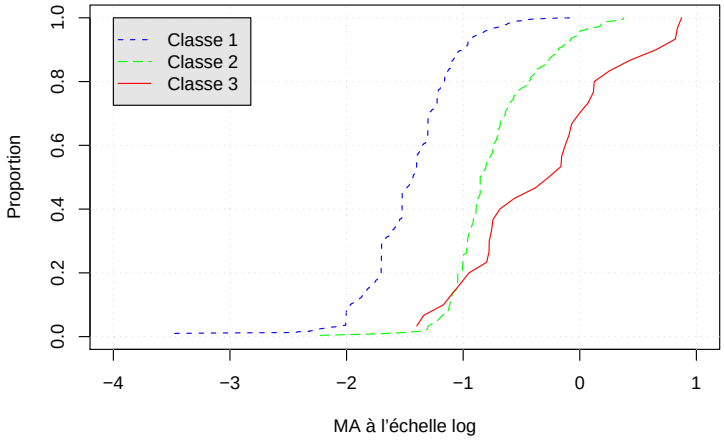

FIG. 6.3 – Fonction de répartition des MA des trois classes.

Classe	Classe 1	Classe 2	Classe 3
Nb de personnes	783 (74, 3%)	241 (22, 9%)	30 (2, 8%)
Moyenne des MA en μT (valeur min ; valeur max)	0,052 (0 ; 0,9)	0,285 (0,01 ;2,38)	1,375 (0,04 ; 7,46)
Moyenne des MG en μT (valeur min ; valeur max)	0,014 (0 ; 0,06)	0,083 (0 ; 0,35)	0,163 (0 ; 1,16)

TAB. 6.3 – Expositions moyennes des adultes de chaque classe en considérant les CM sur 24 heures.

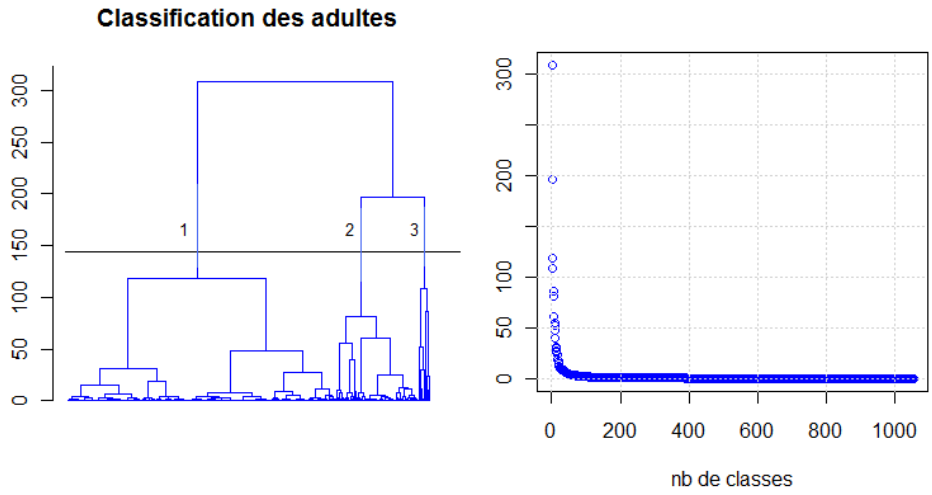

FIG. 6.4 – Dendrogramme de la classification des adultes (figure de gauche) et décroissance de la variance interclasses (figure de droite).

Pour identifier les facteurs pouvant influencer la probabilité d'appartenir aux classes des plus exposés, nous fusionnons les classes 2 et 3 et nous appliquons la régression logistique. Les résultats sont donnés dans le tableau 6.4. Ils montrent que la probabilité d'être classé dans les classes 2 ou 3 croît avec le temps passé dans les transports ferroviaires, dans les supermarchés et la densité de population du département de résidence. Elle est cinq fois plus élevée chez les adultes qui ont leurs foyers proches des lignes aériennes à haute tension (OR=5,1 ; IC=[2,1 ; 12,0]) par rapport à ceux qui habitent loin de ces lignes. Elle est respectivement trois fois, deux fois et 1,7 fois plus importante chez ceux qui ont posé l'EMDEX à côté du radio-réveil (OR=3,2 ; IC=[2,3 ; 4,5]), ceux qui habitent à côté des réseaux ferrés électrifiés (OR=2,0 ; IC=[1,2 ; 3,3]) et ceux qui vivent dans une ville de plus de 2 000 habitants (OR=1,7 ; IC=[1,2 ; 2,5]) par rapport à ceux qui ont éloigné l'EMDEX du radio-réveil, ceux qui ont leurs foyers loin des réseaux ferrés électrifié et qui habitent dans des villes de moins de 2 000 habitants.

Variable	Estimateur	Écart-type	p-value
Temps passé dans les transports ferroviaires	0,36	0,16	0,027
Temps passé dans centres commerciaux	0,33	0,07	$< 0,001$
Densité du département	0,27	0,05	$< 0,001$
Radio-réveil (Oui)	1,17	0,16	$< 0,001$
Population ($>$ 2 000 habitants)	0,55	0,19	0,005
Lignes aériennes à haute tension (Oui)	1,62	0,44	$< 0,001$
Réseaux ferrés électrifié (Oui)	0,69	0,26	0,008

TAB. 6.4 – Variables retenues comme significatives pour la modélisation de la probabilité d'appartenir aux classes des plus exposés pour les adultes en considérant les CM sur 24 heures.

6.5.2 Classification selon les CM enregistrés hors sommeil

6.5.2.1 Les enfants

Le nombre de classes retenues est de 3. Les expositions moyennes enregistrées dans ces classes, en termes de moyennes arithmétique et géométrique sont données dans le tableau 6.5. Les moyennes les plus élevées sont observées dans la classe 3 qui est formée de 57 enfants ($5,9\%$). La moyenne des MA de cette classe est six fois plus élevée que celle de la classe 2 et celle-ci est 3 fois plus importante que celle de la classe 1.

Classe	Classe 1	Classe 2	Classe 3
Nb de personnes	461 ($47,1\%$)	460 ($47,0\%$)	57 ($5,9\%$)
Moyenne des MA en μT (valeur min ; valeur max)	0,016 (0 ; 0,06)	0,055 (0,01 ; 0,21)	0,301 (0,04 ; 1,82)
Moyenne des MG en μT (valeurs min ; valeur max)	0,006 (0 ;0,02)	0,022 (0 ; 0,09)	0,108 (0 ; 0,53)

TAB. 6.5 – Expositions moyennes des enfants de chaque classe en considérant les CM hors sommeil.

Les résultats du test de Moses montrent que les paramètres de dispersion des MA et MG sont différents dans les trois classes prises deux à deux pour les deux moyennes (les bornes supérieures des intervalles de confiance des p-values observées pour 100 réalisations sont inférieures à 0,001). Ces résultats montrent qu'on peut réaliser le test de Wilcoxon-Mann-Whitney pour comparer les paramètres de localisation deux à deux. Les résultats montrent que les enfants de la classe 1 sont globalement moins exposés que

ceux de la classe 2 et ces derniers sont, à leur tour, moins exposés que ceux de la classe 3 pour les MA et les MG (en privilégiant des MA et MG moins élevées dans la classe 1 que dans la classe 2 puis dans la classe 2 que dans la classe 3 dans l'hypothèse alternative, nous obtenons des p-values inférieures à 0,001 pour les deux moyennes). La figure 6.5 donne la position des fonctions de répartition empirique des MA. Elle montre que les écarts entre ces fonctions sont grands. Autrement dit les quantiles des MA de ces classes associés à une proportion donnée sont significativement différents d'une classe à l'autre. Ceci est totalement différent des résultats obtenus en considérant les CM sur 24 heures.

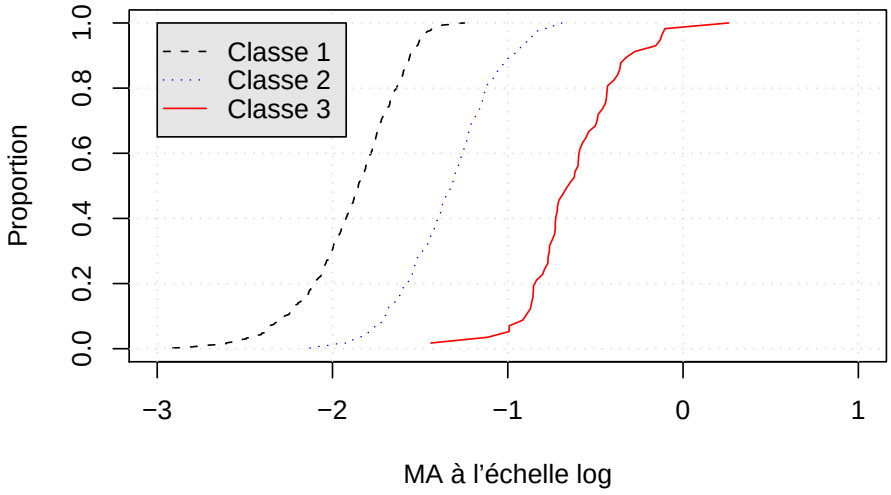

FIG. 6.5 – Fonctions de répartition empirique des MA observées dans les trois classes retenues pour les enfants en considérant les CM hors sommeil.

Pour identifier les facteurs caractérisant la probabilité d'appartenir aux classes des plus exposés, nous appliquons la régression logistique. Nous commençons à chercher les variables caractérisant la probabilité d'appartenir à la classe 3 et nous reprendrons la même démarche en fusionnant les classes 2 et 3.

Les variables expliquant la probabilité d'appartenir dans la classe 3 sont données dans le tableau 6.6. Ce tableau montre que cette probabilité augmente avec le temps passé dans les transports ferroviaires. Elle est plus élevée pour les enfants qui habitent à côté des lignes aériennes à haute tension (OR=3,8 ; IC=[1,1 ; 13,8]) ou à proximité des réseaux ferrés électrifiés

(OR=3,0 ; IC=[1,5 ; 6,1]). Par contre elle diminue avec le temps passé à l'école. Un enfant résidant à côté de ces lignes a trois à quatre fois plus de probabilité d'être classé dans la classe 3 que dans l'une des deux autres.

Variable	Estimateur	Écart-type	p-value
Temps passé dans les transports ferroviaires	1,32	0,66	0,011
Temps passé à l'école	-0,11	0,04	0,005
Lignes aériennes à haute tension (Oui)	1,33	0,66	0,043
Réseaux ferrés électrifiés (Oui)	1,09	0,37	0,003

TAB. 6.6 – Variables retenues comme significatives pour la modélisation de la probabilité d'appartenir à la classe des plus exposés (classe 3) pour les enfants en considérant les CM hors sommeil.

Les résultats de l'estimation du modèle obtenu en considérant les classes 2 et 3 sont donnés dans le tableau 6.7. Ces résultats montrent qu'en fusionnant les classes 2 et 3, on perd l'information portée par les transports ferroviaires, les lignes à haute tension et les réseaux ferroviaires. Cette fois, la probabilité augmente avec la densité du département, le temps passé à regarder ou à jouer devant la télévision et le temps passé sur l'ordinateur. Des différences sont aussi observées selon que l'enfant habite dans une ville de plus ou moin de 2 000 habitant et dans un appartement (immeuble) ou dans un pavillon. Un enfant habitant dans une ville de plus de 2 000 habitants a 1,5 fois plus de probabilité d'être classé dans l'une de deux classes des plus exposés que dans la classe des moins exposés (OR=1,5 ; IC=[1,1 ; 2,0]). S'il habite dans un appartement, il est associé à un odds ratio de 1,8 (IC=[1,1 ; 2,8]).

Variable	Estimateur	Écart-type	p-value
Densité de la population du département	0,92	0,13	$< 0,001$
Temps passé devant la télévision	0,13	0,05	0,012
Temps passé sur l'ordinateur	0,20	0,07	0,002
Habitation (Appartement)	0,58	0,23	0,012
Population ($>$ 2 000 habitants)	0,39	0,16	0,014

TAB. 6.7 – Variables retenues comme significatives pour la modélisation de la probabilité d'appartenir aux classes 2 ou 3 pour les enfants en considérant les CM hors sommeil.

Le tableau 6.7 montre aussi qu'en fusionnant les classes 2 et 3, on perd une information relative à la classe 3 notamment le temps passé dans les transports ferroviaires et la présence des ouvrages électriques à proximité du

foyer. Cela montre bien que la classe 3, des plus exposés est principalement composée par des personnes ayant emprunté les transports ferroviaires ou qui habitent à proximité de ces ouvrages.

6.5.2.2 Les adultes

Nous avons retenu 3 classes d'exposition. Les moyennes des MA et MG observées dans ces classes sont données dans le tableau 6.8. Ce tableau montre que les moyennes calculées sur les individus de la classe 1 (composée de $45,7\%$ des adultes) sont largement inférieures à celles observées dans la classe 2. Les moyennes les plus élevées sont enregistrées dans la classe 3 (0,619 μT pour les MA et 0,124 μT pour les MG) composée de $5,0\%$ des adultes.

Classe	Classe 1	Classe 2	Classe 3
Nb de personnes	482 ($45,7\%$)	519 ($49,3\%$)	53 ($5,0\%$)
Moyenne des MA en μT (valeur min ; valeur max)	0,038 (0 ; 0,24)	0,106 (0 ; 1,35)	0,619 (0,06 ; 7,14)
Moyenne des MG en μT (valeur min ; valeur max)	0,015 (0 ; 0,08)	0,039 (0 ; 0,17)	0,124 (0 ; 0,98)

TAB. 6.8 – Expositions moyennes des adultes de chaque classe en considérant les CM hors sommeil.

Pour voir si ces différences sont statistiquement significatives, nous réalisons des tests de comparaison des paramètres de localisation des MA et des MG des classes. Pour cela, nous avons appliqué le test d'égalité des paramètres de dispersion de Moses. Les résultats du test de Moses montrent que, pour 100 réalisations, les p-values moyennes sont inférieures à $\alpha = 0,05$ et les bornes supérieures des intervalles de confiance sont aussi inférieures à α pour les deux moyennes. Ils indiquent que les paramètres de dispersion ne peuvent pas être considérés comme égaux dans les trois classes. Avec ce résultat, nous appliquons le test de Fligner-Policello pour comparer les paramètres de localisation. Pour avoir un ordre de classement nous réalisons les tests deux à deux. L'hypothèse nulle est "les paramètres de localisation sont identiques pour la moyenne et les classes considérées". Nous privilégions des paramètres de localisation plus élevés dans la classe 2 par rapport à la classe 1 et dans la classe 3 par rapport à la classe 2. Ce choix peut être guidé par la position des fonction de répartition empirique (la figure 6.6 donne celles des MA dans les trois classes). Les résultats des tests rejettent les hypothèses nulles avec des p-values inférieures à 0,001 pour les deux moyennes. Les individus de la classe 1 sont moins exposés que ceux de la classe 2 et ces derniers sont moins exposés que ceux de la classe 3 en termes de MA et MG.

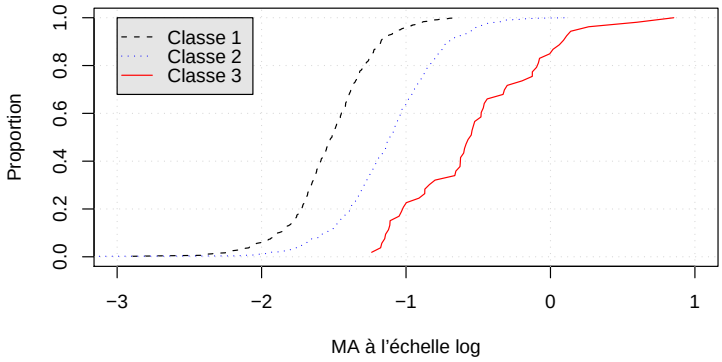

FIG. 6.6 – Fonctions de répartition empirique des MA observées dans les trois classes retenues pour les adultes en considérant les CM hors sommeil.

Les résultats de l'estimation du modèle de régression logistique retenu pour l'identification des variables influençant la probabilité d'appartenir à la classe 3 sont donnés dans le tableau 6.9. Ce tableau montre que plus on passe de temps dans les transports ferroviaires ou dans les centres commerciaux plus le probabilité d'être classé dans la classe des plus exposés (classe 3) augmente. Cette probabilité est 4,8 fois plus élevée pour une personne ayant son foyer de résidence à côté de lignes aériennes à haute tension qu'une autre qui habite dans un foyer éloigné de ces lignes (OR=4,8 ; IC=[1,6 ; 14,7]).

Variable	Estimateur	Écart-type	p-value
Temps passé dans les transports ferroviaires	0,46	0,22	0,039
Temps passé dans les centres commerciaux	0,26	0,10	0,012
Lignes aériennes à haute tension (Oui)	1,56	0,57	0,006

TAB. 6.9 – Estimation du modèle expliquant la probabilité d'appartenir à la classe des plus exposés pour les adultes en considérant les CM hors période de sommeil.

6.6 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la recherche de classes d'exposition. Nous avons retenu trois classes pour chaque type de population en considérant les CM sur 24 heures ou hors sommeil. En termes de moyennes des MA et des MG, on peut dire que la classe 1 est composée des

personnes qui sont faiblement exposées. Les proportions des individus qui la composent dépendent du type de population et du scénario de données considéré. Pour les CM sur 24 heures, elles sont de $31,5\%$ pour les enfants (les expositions moyennes sont de 0,036 μT pour les MA et 0,020 μT pour les MG) et $74,3\%$ pour les adultes (les moyennes des MA et des MG sont respectivement de 0,052 μT et 0,014 μT). Pour les CM hors sommeil, elles sont de $47,1\%$ pour les enfants (les expositions moyennes sont de 0,016 μT pour les MA et 0,006 μT pour les MG) et $45,7\%$ pour les adultes (les moyennes observées sont de 0,038 μT pour les MA et 0,015 μT pour les MG).
Quant à la classe 2, on peut dire qu'elle est formée des individus ayant une exposition normale due à l'omniprésence des sources de CM. Les moyennes observées dans cette classe sont globalement entre 2 et 5 fois plus élevées que dans la classe 1. Les individus les plus exposés sont classés dans la classe 3. Leurs proportions sont faibles ($10,4\%$ et $5,9\%$ pour les enfants et $2,8\%$ et $5,3\%$ pour les adultes respectivement pour les CM sur 24 heures et hors sommeil).

Les tests de comparaison des paramètres de localisation des MA et des MG dans les classes ont montré que les différences entre les classes sont significatives. Les valeurs minimales des MA et des MG ne sont pas forcement différentes d'une classe à l'autre. Comme les classes sont disjointes, cela veut dire que les MA et les MG ne sont pas les seuls indicateurs discriminatoires.

Pour identifier les facteurs caractérisant la probabilité d'être classé dans la classe des plus exposés, nous avons modélisé la probabilité d'appartenir à la classe 3 à l'aide d'une régression logistique.

- Pour les classes obtenues avec les CM sur 24 heures, cette probabilité est plus élevée pour les personnes qui ont posé l'EMDEX à proximité d'un radio-réveil et ou qui ont leurs foyers à côté des réseaux ferrés électrifiés. Pour les adultes, la probabilité augmente avec le temps passé dans les transports ferroviaires, dans les centres commerciaux et avec la densité de population du département de résidence. Elle est aussi plus importante pour les adultes qui habitent dans une ville de plus de 2 000 habitants et qui ont leurs foyers à côté des lignes aériennes à haute tension. Pour les enfants, elle baisse avec le temps passé à l'école et dans les transports non électriques.
- Pour les CM hors sommeil, on trouve comme facteurs favorisant cette probabilité le temps passé dans les transports ferroviaires, le fait d'avoir son foyer à côté des lignes aériennes à haute tension pour les deux populations. Elle croît avec le temps passé dans les centres commerciaux pour les adultes et est plus élevée pour les enfants qui habitent à côté des réseaux ferrés électrifiés. D'autres variables apparaissent lorsqu'on s'intéresse à la probabilité d'appartenir aux classes 2 ou 3 pour les enfants comme le temps passé devant la télévision ou sur ordinateur.

Chapitre 7

Conclusion et perspectives

7.1 Conclusion générale

Dans cette thèse de doctorat, une étude de l'exposition de la population au champ magnétique généré par le transport et l'utilisation d'électricité est réalisée avec des mesures personnelles. C'est la première fois qu'une telle étude est menée à l'échelle d'un pays. Cette thèse est principalement composée de trois points :

1. la sélection des individus et la collecte des informations,
2. l'analyse des expositions moyennes,
3. la caractérisation de ces dernières.

7.1.1 Sélection des individus et collecte des informations

Cette thèse a pour but d'estimer et de caractériser l'exposition de la population française au champ magnétique 50 Hz. Pour cela, un protocole relatif à l'étude a été établi. Dans ce protocole, il a été décidé de réaliser l'étude sur 1 000 enfants de 14 ans et moins et 1 000 adultes de 15 ans et plus. Les individus sont sélectionnés par la méthode de tirage aléatoire. Cette méthode est basée sur la sélection de ménages à partir de numéros de téléphone distribués en France métropolitaine en respectant les proportions de la distribution des ménages dans les 22 régions. Ces numéros de téléphones comprennent les numéros en liste rouge et ceux des ménages ayant seulement un téléphone portable. Cette phase de sélection des individus et de collecte des informations a été réalisée par l'institut de sondage MV2 Conseil. Chaque personne sondée a porté un EMDEX II mesurant et enregistrant toutes les trois secondes le CM auquel elle est exposée pendant une période minimale de 24 heures. Chaque personne a rempli aussi un emploi du temps relatif à la période de mesure. De son côté, l'enquêteur a rempli, avec l'aide du volontaire, un questionnaire relatif à la personne sondée et à l'environnement électromagnétique de son foyer, et a noté les coordonnées GPS à l'entrée du foyer du volontaire. Ces dernières ont permis, avec l'aide d'ERDF et de RTE, d'identifier à proximité du foyer l'existence ou non de réseaux électriques allant de la basse tension à la très haute tension, et de réseaux ferrés électrifiés.

Au total, 95 362 numéros de téléphone ont été composés pour avoir une base de données validée de 2 032 sujets (1 054 adultes et 978 enfants). Les difficultés rencontrées par MV2 Conseil à recruter des enfants et le temps moyen de communication téléphonique pour le recrutement d'un volontaire (70 minutes) sont les deux points marquant de cette campagne de recueil des données. Les parents étaient réticents à participer à l'étude dès que le choix tombait sur un enfant en bas âge. Cela a conduit à une proportion moins élevée des enfants de moins de 6 ans par rapport à la population nationale.

Pour les adultes, on a observé des proportions plus élevées pour la classe d'âge 34-50 ans en comparaison avec la population nationale. Cependant l'objectif initial de 1 000 enfants et 1 000 adultes a été atteint.

7.1.2 Estimation des expositions moyennes

Le premier objectif de l'étude est l'estimation de l'exposition d'un échantillon représentatif de la population. Pour cela, les moyennes arithmétique et géométrique sont calculées pour chaque série de CM. Les expositions moyennes sont ensuite estimées par les moyennes des MA et des MG. Elles sont, pour les enfants, de 0,09 et 0,02 μT respectivement pour les MA et les MG et 0,14 et 0,03 μT pour les adultes. Le Centre international de recherche sur le cancer (CIRC) a, sur la base des résultats des méta-analyses, classé le champ magnétique d'extrêmement basse fréquence comme cancérogène possible (groupe 2B) pour le risque de leucémie de l'enfant en rapport avec des expositions élevées et prolongées, supérieures à 0,4 μT en moyenne sur 24 heures, sans qu'un lien de causalité n'ait été demontré.
Dans cette étude, $3,1\%$ des enfants (30 enfants) ont observé une MA supérieure à 0,4 μT. Cette proportion nous a paru élevée par rapport à celle attendue en se référant à la littérature. Elle est en particulier supérieure à celle observée dans la méta-analyse d'Ahbom ($1,06\%$ au total, $0,41\%$ en Europe et $2,64\%$ en Amerique) mais inférieure aux proportions observées par Grenn et McBride au Canada. Pour ces deux dernières études, les proportions des enfants ayant observé une moyenne supérieure à $0,4\mu$T sont respectivement de $3,65\%$ et $4,27\%$. Nous avons cherché à expliquer ces valeurs élevées, ce qui rentre dans le deuxième objectif de l'étude, qui est l'identification des sources de champ magnétique.

Cette analyse a montré que les principales sources de CM mesuré par les EMDEX portés par ces 30 enfants sont dans 80% des cas des radio-réveils. En fait, 24 d'entre eux ont déclaré avoir posé l'EMDEX à moins de 50 cm du radio-réveil pendant la nuit. Cette information est validée en examinant les données de mesure de CM couplées aux emplois du temps. Mais cette vérification ne permet pas de dire si les CM mesurés à moins de 50 cm des radio-réveils reflètent ou non l'exposition de la personne. C'est pourquoi l'analyse des données a été réalisée avec deux scénarios : en considérant les CM mesurés sur toute la période de mesure et hors période de sommeil.
En considérant les CM mesurés hors période de sommeil, nous avons éliminé les CM générés par les radio-réveils soit, en moyenne, 8 heures d'observation. Les expositions moyennes calculées sur ce nouveau scénario sont, pour les enfants, de 0,05 et 0,02 μT respectivement pour les MA et les MG et de 0,10 et 0,03 μT pour les adultes. Au total, $1,1\%$ des enfants ont, dans scénario, observé une MA supérieure à 0,4 μT, ce qui est plus conforme aux données de la littérature.

Si l'exposition due aux radio-réveils reflète l'exposition de la personne, alors les études épidémiologiques, qui considèrent pour la plupart uniquement le champ magnétique généré par les lignes à haute tension, comportent un large biais de classification entre les exposés et les non exposés. Si, comme nous le croyons, cette source ne reflète pas l'exposition personnelle, alors environ 1% des enfants sont soumis à une exposition moyenne supérieure à 0,4 μT sur 24 heures.

Pour étudier les sources de champ magnétique tels que les réseaux à haute tension et les réseaux ferrés électrifiés, des test de comparaison des MA et des MG observées au domicile et sur 24 heures ont été réalisés. Les résultats ont montré qu'il n'y pas de différences entre les moyennes observées par les individus habitant à proximité de ces deux types de réseaux. En revanche, comme attendu, l'exposition est moins élevée pour les personnes habitant loin de ces ouvrages électriques par rapport à ceux habitant à côté de ces derniers. D'autres tests de comparaison ont été aussi réalisés. Ils ont montré que les enfants sont moins exposés que les adultes et que l'exposition est plus élevée en Ile-de-France que dans les autres régions. Ces tests ont également montré que l'exposition est plus élevée au domicile qu'à l'extérieur pour les enfants alors qu'ils ont indiqué le contraire pour les adultes.

7.1.3 Caractérisation des expositions

La phase de caractérisation des expositions est réalisée de deux manières.

1. **Caractérisation des expositions moyennes.**
   Des modèles de régression ont été réalisés entre les moyennes arithmétiques et géométriques sur 24 heures et hors période de sommeil d'une part et les informations recueillies dans les emplois du temps et les questionnaires d'autre part. Ces modèles ont permis d'identifier des facteurs favorisant une exposition moyenne plus élevée comme le fait d'habiter à proximité des lignes aériennes à haute tension, des réseaux ferrés électrifiés, dans une ville de plus de 2 000 habitants et dans un appartement. Les expositions moyennes augmentent aussi avec le temps passé dans les transports ferroviaires, sur ordinateur et la densité de population du département. D'autre facteurs apparaissent selon le scénario et le type de moyenne considérés comme les radio-réveils pour les moyennes su 24 heures et le temps passé dans les centres commerciaux pour les adultes. Les taux de variance expliquée par les facteurs identifiés ne dépassent pas les 30% : les modèles obtenus ne sont pas prédictifs. Cela permet de conclure que ces facteurs ne permettent pas d'expliquer, à eux seuls, les expositions moyennes. Autrement dit, d'autres informations sont nécessaires pour bien caractériser les expositions moyennes.

2. **Recherche de classes d'exposition et caractérisation de la classe des plus exposés.**
   Pour répartir l'ensemble des individus en plusieurs classes d'exposition, chaque série de CM est décrite par 7 indicateurs (la valeur maximale, la MA, la MG, la médiane, l'écart-type et le RCMS). Une classification ascendante hiérarchique est ensuite appliquée sur ces descripteurs centrés et réduits. Pour chaque scénario (CM sur 24 heures et hors période de sommeil), nous avons retenu trois classes d'exposition (une classe des moins exposés - classe 1, une classe des moyennement exposés - classe 2 et une classe des plus exposés - classe 3). La classe 3 est la moins représentée. Elle compte entre 2,8 et $10,4\%$ des individus selon le scénario et le type de population considérés. Pour identifier les facteurs favorisant la probabilité d'appartenir aux classes 2 et 3, nous avons modélisé cette probabilité à l'aide d'une régression logistique. Les variables explicatives testées dans cette régression sont celles utilisées pour caractériser les expositions moyennes. Les résultats de cette régression dépendent du scénario considéré. Pour les CM sur 24 heures, la probabilité d'appartenir aux classes 2 et 3 est plus élevée chez les personnes habitant à proximité des réseaux ferrés électrifiés ou ayant posé l'EMDEX à côté du radio-réveil pour les deux populations. Elle est aussi plus élevée chez les adultes habitant à proximité des lignes aériennes à haute tension ou dans les villes de plus de 2 000 habitants. Elle augmente également avec la densité de population du département, le temps passé dans les centres commerciaux et dans les transports ferroviaires.
   Pour les CM hors sommeil, la probabilité d'appartenir aux classes 2 et 3 est plus élevée chez les personnes habitant à proximité des lignes aériennes à haute ou très haute tension et augmente avec le temps passé dans les transports ferroviaires pour les deux populations. D'autres variables apparaissent selon le type de population considéré comme habiter à proximité des réseaux ferrés électrifiés pour les enfants ou le temps passé dans les centres commerciaux pour les adultes.

7.2 Perspectives

L'analyse des expositions moyennes a montré que les variables retenues ne permettent pas à elles seules de caractériser ces moyennes. Dans la suite, la base de données sera complétée en incluant la présence ou non de lignes à basse et moyenne tension et de postes de transformation à proximité des domiciles des volontaires. On améliorera également l'indicateur d'exposition au champ généré par les réseaux électriques en prenant en compte le courant si possible. Ces informations pourraient améliorer la caractérisation des MA et des MG en termes de variance expliquée.

La base de données est quant à elle riche en information. Elle constitue en fait la première base de données avec des mesures personnelles à l'échelle de la France. Elle pourrait servir à valider les modèles physiques d'estimation du champ magnétique 50 Hz.

7.3 Liste des publications

La méthodologie de cette étude a fait l'objet d'un article soumis dans la révue « Radiation Protection Dosimetry »
Methodology of a study on the French population exposure to 50 Hz magnetic fields.
M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], A. Carlsberg[6].

Les résultats seront soumis dans la révue Bioelectomagnetics. Les travaux menés dans le cadre de cette thèse ont été ou seront présentés dans les congrès suivants :

1. Exposition de la population française au champ magnétique 50 Hz
   15ème Colloque International et Exposition sur la Compatibilité Electromagnétique, 7-9 avril 2010, Limoge, France.
   M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5],A. Carlsberg[6].
2. Analyse des données de réseaux électriques dans l'étude EXPERS
   15ème Colloque International et Exposition sur la Compatibilité Electromagnétique, 7-9 avril 2010, Limoge, France.
   I. Magne[1], M. Bédja[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], M. Le Lay[5], A. Carlsberg[6], J-L. Richard[7].
3. Exposure of the French population to 50 Hz magnetic fields : EXPERS study.
   Third European IRPA Congress, 14-18 june 2010, Helsinki, Finland.
   M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], A. Carlsberg[6].
4. Analysis of electric network data in the EXPERS study
   Third European IRPA Congress, 14-18 june 2010, Helsinki, Finland.
   I. Magne[1], M. Bédja[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], M. Le Lay[5], A. Carlsberg[6], J-L. Richard[7].
5. French population exposure to 50 Hz magnetic fields : EXPERS study.
   32th Annual Meeting The Bioelectromagnetics Society, June 14-18 2010, Seoul, Korea.

M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], A. Carlsberg[6].

6. Analysis of electric network data in the EXPERS study.
   32$^{\text{th}}$ Annual Meeting The Bioelectromagnetics Society, June 14-18 june 2010, Seoul, Korea.
   I. Magne[1], M. Bédja[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], M. Le Lay[5], A. Carlsberg[6], J-L. Richard[7].

D'autres présentations ont eu lieu pendant la thèse :

1. French population exposure to 50 Hz magnetic fields : first results in île-de-France and Rhône-Alpes regions.
   The 36$^{\text{th}}$ annual meeting of the European Radiation Research society (ERR), September 1-4, 2008, Tours, France.
   M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G.Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]
2. Methodologie of the 50 Hz magnetic fields exposure of the French population.
   The 36th annual meeting of the European Radiation Research society (ERR), September 1-4, 2008, Tours, France.
   M. Bédja[1],I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]
3. French population exposure to 50 Hz magnetic fields : first results in île-de-France (IDF) and Rhône-Alpes (RHA) regions.
   International Workshop on Biological Effects of Electromagnetic Fields, September 28$^{\text{th}}$-October 2$^{\text{nd}}$ 2008, Palermo, Italia.
   M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]
4. French population exposure to 50 Hz magnetic fields : intermediate results.
   International Colloquium on « Power Frequency Electromagnetic Fields-ELF/EMF », 3-4 June 2009, Sarajevo, Bosnia Herzegovina.
   M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]
5. French population exposure to 50 Hz magnetic fields : intermediate results.
   The Joint Meeting of The Bioelectromagnetics Society and The European Bioelectromagnetics Association, 14-19 June 2009, Davos, Switzerland.
   M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]
6. Exposition de la population française au champ magnétique 50 Hz : résultats partiels.

Société Française de Radio Protection (SFRP). Congrès National de Radioprotection, 16-18 juin 2009, Angers, France.
M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]

7. Exposition de la population française au champ magnétique 50 Hz : premiers résultats pour les régions Île-de-France (IDF) et Rhône-Alpes (RHA).
Société Française de Radio Protection (SFRP) : journée Rayons Non-Ionisants (RNI), 7 octobre 2008, Grenoble, France.
M. Bédja[1], I. Magne[1], M. Souques[2], J. Lambrozo[2], L. Le Brusquet[3], G. Fleury[3], A. Azoulay[4], F. Deschamps[5], S. Ruszczynski[6]

1-Laboratoire des Matériels Electriques, EDF, Moret-sur-Loing, France.
2-Service des Etudes Médicales, EDF Gaz de France, Paris, France.
3-Département Signaux et Systèmes Electroniques, SUPELEC, Gif-sur-Yvette, France.
4-Département Électromagnétisme, SUPÉLEC, Gif-sur-Yvette, France.
5-Centre National d'Expertise Réseaux, RTE, La Défense, France.
6-Département santé, MV2 Conseil, Montrouge, France.
7-Electricité Réseau Distribution France (ERDF), La Défense, France.

Annexe 1 : Compléments du chapitre 4

Preuve du théorème 4.1

On note :
W_{XY} le nombre de couples (X_i, Y_j) tel que $X_i < Y_j$ et $H_{ij} = 1_{(X_i<Y_j)}$.

Sous H_0, H_{ij} est une variable aléatoire qui suit une loi de Bernoulli de paramètre $\frac{1}{2}$ et

$$W_{XY} = \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} X_i < Y_j$$

$Y_{(1)}$ a pour rang r_1 : $r_1 - 1$ valeurs de X sont inférieures à $Y_{(1)}$
$Y_{(2)}$ a pour rang r_2 : $r_2 - 2$ valeurs de X sont inférieures à $Y_{(2)}$
...

$\sum_{i=1}^{n_2} \sum_{j=1}^{n_1} H_{ij} = \sum_{j=1}^{n_1} (r_j - j) = W_s - \frac{1}{2} n_1(n_1 + 1)$ et

$W_{XY} = \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} H_{ij}$

Sous H_0, W_{XY} est une somme de $n_1 \times n_2$ variables aléatoires (non indépendantes) de Bernoulli de paramètre $\frac{1}{2}$.

a) Calcul d'espérances

$\mathbb{E}(W_{XY}) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} \mathbb{E}(H_{ij}) = \frac{1}{2} n_1 n_2$ car $\mathbb{E}(W_{XY}) = P(X_i < Y_j) = \frac{1}{2}$

Par conséquent,

$\mathbb{E}(W_s) = \mathbb{E}(W_{XY}) + \frac{1}{2} n_1(n_1 + 1) = \frac{1}{2} n_1 n_2 + \frac{1}{2} n_1(n_1 + 1) = \frac{1}{2} n_1(n_1 + n_2 + 1)$.

b) Calcul de la variance

$$\mathbb{V}(W_{XY}) = \mathbb{V}\left(\sum_{i=1}^{n_2} \sum_{j=1}^{n_1} H_{ij} \right)$$

$\mathbb{V}(W_{XY}) = \sum_{(i,j)=(i',j')} cov(H_{ij}, H_{i'j'}) + \sum_{(i,j)\neq(i',j')} cov(H_{ij}, H_{i'j'})$ et

$\sum_{(i,j)=(i',j')} cov(H_{ij}, H_{i'j'}) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} \mathbb{V}(H_{ij}) = \frac{1}{4} n_1 n_2$ car sous H_0,

$cov(H_{ij}, H_{ij}) = \mathbb{V}(H_{ij}) = \frac{1}{4}$.

c) Calcul des covariances
Par définition de la covariance, pour tout couple (i, j) et (i', j') :

$cov(H_{ij}, H_{i'j'}) = \mathbb{E}(H_{ij}H_{i'j'}) - \mathbb{E}(H_{ij})\mathbb{E}(H_{i'j'}) = \mathbb{E}(H_{ij}H_{i'j'}) - \frac{1}{4}$

On distingue trois possibilités :

1. $i \neq i'$ et $j \neq j'$.
2. $i = i'$ et $j \neq j'$.
3. $i \neq i'$ et $j = j'$.

Cas 1 : $i \neq i'$ et $j \neq j'$
Dans ce premier cas, on a :

$\mathbb{E}(H_{ij}H_{i'j'}) = P(X_i < Y_i, X_{i'} < Y_{j'}) = \frac{1}{4}$ et $cov(H_{ij}, H_{i'j'}) = \frac{1}{4} - \frac{1}{4} = 0.$

Cas 2 : $i = i'$ et $j \neq j'$
Dans ce second cas, on a :
$cov(H_{ij}, H_{ij'}) = P(X_i < Y_i, X_i < Y'_j) = \frac{1}{3}$ et $cov(H_{ij}, H_{ij'}) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

Cas 3 : $i \neq i'$ et $j = j'$

$cov(H_{ij}, H_{i'j}) = P(X_i < Y_i, X'_i < Y_j) = \frac{1}{3}$ et $cov(H_{ij}, H_{i'j}) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

Pour finir, il y a :
- n_1 façons de choisir j
- $n_1 - 1$ façons de choisir j'
- n_2 façons de choisir i
- $n_2 - 1$ façons de choisir i'

En rassemblant tous les termes, on trouve :

$\mathbb{V}(W_{XY}) = \frac{n_1 n_2}{4} + \left\{ \frac{n_2 n_1 (n_1 - 1)}{12} + \frac{n_1 n_2 (n_2 - 1)}{12} \right\}$

$\mathbb{V}(W_{XY}) = \frac{1}{12} n_1 n_2 \left(3 + n_1 - 1 + n_2 - 1\right)$

$\mathbb{V}(W_{XY}) = \frac{1}{12} n_1 n_2 \left(n_1 + n_2 + 1\right)$ et $\mathbb{V}(W_s) = \frac{1}{12} n_1 n_2 \left(n_1 + n_2 + 1\right).$

Preuve du théorème de la loi faible des grands nombres

La preuve de ce théorème fait appel à l'inégalité de Bienaymé-Tchebicheff :

Proposition 1 *: Inégalité de Bienaymé-Tchebicheff*
Soit X une variable aléatoire de variance finie σ^2 et soit $\epsilon > 0$.
Alors, on a :

$$P\left(|X| \geq \epsilon\right) \leq \frac{\mathbb{E}\left(X^2\right)}{\epsilon^2} \tag{7.1}$$

$$P\left(|X - \mathbb{E}(X)| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2} \tag{7.2}$$

Preuve

Pour démontrer les inégalités (7.2) et (7.2), on va supposer que la variable X est discrète. Dans le cas où elle est continue, il suffit de remplacer la somme par l'intégrale.

En utilisant la définition de la probabilité d'un événement A, défini dans un espace probabilisé (Ω, w, p_w) et celle de l'espérance mathématique, on a :

$\mathbb{E}\left(X^2\right) = \sum_\omega p_\omega X(\omega)^2 \geq \epsilon^2 \sum_{\omega:|X(\omega)|\geq\epsilon} p_\omega = \epsilon^2 P\left(|X| \geq \epsilon\right)$ où p_w est la probabilité de l'évennement w dans l'espace Ω

$\mathbb{E}\left(X^2\right) \geq \epsilon^2 P\left(|X| \geq \epsilon\right)$ ou encore $P\left(|X| \geq \epsilon\right) \leq \frac{\mathbb{E}\left(X^2\right)}{\epsilon^2}$.

Pour démontrer l'autre inégalité, il suffit d'appliquer la variable $X - \mathbb{E}(X)$ à la première inégalité.

Preuve du théorème de la loi faible des grands nombres

Les variables aléatoires X_i sont indépendantes et identiquement distribuées (iid) :

$\mathbb{E}\left(\bar{X}_n\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{1}{n}(n\mu) = \mu$

$\mathbb{V}\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{1}{n}\sigma^2$

Appliquons l'inégalité de Bienaymé-Tchebicheff à la variable $\bar{X}_n$.
Il vient :

$$P\left(\left|\bar{X}_n - \mu\right| \geq \frac{t\sigma}{\sqrt{n}}\right) \leq \frac{1}{t^2}.$$

En posant $\epsilon = \frac{t\sigma}{\sqrt{n}}$, on a $t = \frac{\epsilon\sqrt{n}}{\sigma}$, d'où $\frac{1}{t^2} = \frac{\sigma^2}{n\epsilon^2}$. Alors :

$$P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

L'événèment $\left|\bar{X}_n - \mu\right| \leq \epsilon$ est l'événèment contraire de $\left|\bar{X}_n - \mu\right| > \epsilon$, donc

$P\left(\left|\bar{X}_n - \mu\right| \leq \epsilon\right) = 1 - P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right)$.

Puisque $P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$, on a $1 - P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$, d'où finalement :

$$P\left(\left|\bar{X}_n - \mu\right| \leq \epsilon\right) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

Du fait que toute probabilité est inférieure à 1, on a l'encadrement :

$$1 - \frac{\sigma^2}{n\epsilon^2} \leq P\left(\left|\bar{X}_n - \mu\right| \leq \epsilon\right) \leq 1$$

En utilisant le théorème de Gendarme, on en déduit immédiatement que :

$$\forall \epsilon > 0, \lim_{n \to +\infty} P\left(\left|\bar{X}_n - \mu\right| \leq \epsilon\right) = 1$$

Preuve du théorème central limite

Une des manières existantes pour démontrer le théorème central limite est d'utiliser la fonction caractéristique de la variable Z_n. Pour cela, on définit en quelques lignes la fonction caractéristique d'une variable aléatoire.

Fonction caractéristique d'une variable aléatoire

Dans ce paragraphe, nous introduisons la notion de fonction caractéristique d'une variable aléatoire, un outil important en calcul de probabilité, et qui s'apparente à la transformée de Fourier en traitement de signal.

On notera $< x, y >$ le produit scalaire de deux vecteurs de $\mathbb{R}^n$. La fonction complexe $x \longmapsto e^{j<x,y>}$ ($j^2 = -1$) est continue de module 1. Donc si X est une variable aléatoire à valeurs dans $\mathbb{R}^n$ et $u \in \mathbb{R}^n$, on peut considérer que $e^{j<u,X>}$ est une variable aléatoire à valeurs complexes (ses parties réelle et imaginaire qui sont aussi des variables aléatoires sont respectivement $Y = \cos(< u, X >)$ et $Z = \sin(< u, X >)$). Ces variables aléatoires réelles sont bornées par 1, elles admettent donc une espérance. On définit donc l'espérance de $e^{j<u,X>}$ par :

$$\begin{aligned} \mathbb{E}(e^{j<u,X>}) &= \mathbb{E}(Y) + j\mathbb{E}(Z) \\ \mathbb{E}(e^{j<u,X>}) &= \mathbb{E}\left(\cos(< u, X >)\right) + j\mathbb{E}\left(\sin(< u, X >)\right) \end{aligned}$$

Définition 2 *Si X est une variable aléatoire à valeurs dans $\mathbb{R}^n$, sa fonction caractéristique est la fonction ϕ_X de $\mathbb{R}^n$ dans $\mathbb{C}$ définie par :*

$$\phi_X(u) = \mathbb{E}\left(e^{j<u,X>}\right) \tag{7.3}$$

Cette fonction ne dépend que de la loi de X.

La fonction caractéristique d'une variable aléatoire à des propriétés importantes sur la théorie des probabilités mais nous n'allons nous intéresser qu'à celles nous permettant de démontrer le théorème de la limite centrale.

Exemple : Fonction caractéristique de la loi normale centrée et réduite
Soit X une variable aléatoire définie dans $\mathbb{R}$ de loi normale centrée et réduite. Calculons sa fonction caractéristique notée φ_X.
Par définition,

$$\varphi_X(u) = \mathbb{E}\left(e^{juX}\right)$$

$$\varphi_X(u) = \int_{-\infty}^{+\infty} e^{jux} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx.$$

En dérivant à l'intérieur de l'intégrale, on a :

$$\varphi'_X(u) = j \int_{-\infty}^{+\infty} x e^{jux} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

En posant $U'(x) = xe^{-\frac{1}{2}x^2}$ et $V(x) = e^{jux}$ puis en intégrant par parties, on trouve :

$$\varphi'_X(u) = -u \int_{-\infty}^{+\infty} e^{jux} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \text{ ou encore}$$

$$\varphi'_X(u) = -u\varphi_X(u)$$

C'est une équation différentielle de premier ordre dont on connaît la solution.

$\varphi_X(u) = e^{-\frac{1}{2}u^2}$ car $\varphi_X(0) = 1$.
La fonction caractéristique φ de la loi normale centrée et réduite est définie pour tout $u \in \mathbb{R}$ par :

$$\varphi(u) = e^{-\frac{1}{2}u^2}$$

Proposition 3 *Si X et Y sont des variables aléatoires indépendantes à valeurs dans $\mathbb{R}^n$, de fonctions caractéristiques ϕ_X et ϕ_Y, alors la fonction caractéristique de $X + Y$ est donnée par (7.4).*

$$\phi_{X+Y} = \phi_X \phi_Y \tag{7.4}$$

Preuve
Par définition, $\phi_X(u) = \mathbb{E}\left(e^{j<u,X>}\right)$ et donc

$$\phi_{X+Y}(u) = \mathbb{E}\left(e^{j<u,X+Y>}\right)$$

$$\phi_{X+Y}(u) = \mathbb{E}(e^{j(<u,X>+<u,Y>)}) = \mathbb{E}(e^{j<u,X>} e^{j<u,Y>})$$

Comme les variables X et Y sont indépendantes,

$$\mathbb{E}\left(e^{j<u,X>} e^{j<u,Y>}\right) = \mathbb{E}(e^{j<u,X>})\mathbb{E}(e^{j<u,Y>}) = \phi_X(u)\phi_Y(u)$$

Preuve du théorème central limite

Notons $Y_i = \frac{X_i - \mu}{\sigma}$, alors $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$.
Les Y_i sont centrées et réduites.

Lemme 4 *La fonction caractéristique ϕ_{Y_i} admet un développement limité d'ordre 2 en 0 donné, pour tout réel t par (7.5).*

$$\phi_{Y_i}(t) = 1 - \frac{t^2}{2} + o(t^2) \tag{7.5}$$

En fait la formule de Taylor avec reste intégrale écrite à l'ordre 2 donne, pour tout réel x :

$e^{jx} = 1 + jx - \int_0^x (x-u)e^{ju}du$

Après un changement de variable, on a

$e^{jx} = 1 + jx - x^2 \int_0^1 (1-u)e^{jux}du$

soit, puisque

$\int_0^1 (1-u)du = \frac{1}{2}$

$e^{jx} - (1 + jx - \frac{x^2}{2}) = -x^2 \int_0^1 (1-u)(e^{jux} - 1)du$

$|e^{jx} - (1 + jx - \frac{x^2}{2})| \leq x^2 \int_0^1 |(1-u)(e^{jux} - 1)|du \leq x^2$
Ainsi

$e^{jx} = 1 + jx - \frac{x^2}{2} + o(x^2)$ et $\phi_{Y_i}(t) = 1 - \frac{t^2}{2} + o(t^2)$ car $\mathbb{E}(Y_i) = 0$ et $\mathbb{E}(Y_i^2) = 1$

La fonction caractéristique de Z_n est donc définie par :

$$\phi_{Z_n}(t) = \prod_{i=1}^{n} \phi_{Y_i}(\tfrac{t}{\sqrt{n}})$$

En utilisant(7.5),

$$\phi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n$$

$$\lim_{n\to+\infty} \phi_{Z_n}(t) = \lim_{n\to+\infty} \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n = e^{-\frac{1}{2}t^2}$$

La suite des fonctions caractéristiques $(\phi_{Z_n})_{n\in\mathbb{N}}$ converge simplement vers la fonction caractéristique de la loi normale centrée et réduite qui est continue

en 0. En utilisant le théorème de continuité de Lévy, qui affirme que la convergence des fonctions caractéristiques implique la convergence en loi [33, 46], on en déduit que la suite des variables $(Z_n)_{n\in\mathbb{N}}$ converge vers la loi normale centrée et réduite c'est-à-dire :

$$\lim_{n\to+\infty} P(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

ANNEXE 2 : Lettre de la direction générale de la santé

ANNEXE 3 : Emploi du temps

ANNEXE 4 : Questionnaire

Bibliographie

[1] Wertheimer N., Leeper E., Electrical wiring configurations childhood cancer, American Journal of Epidemiology 109 :273-284, 1979.

[2] World Health Organization, Extremely low frequency fields. Environmental Health Criteria 238, Genève, World Health Organization, 2007.

[3] International Agency for Research on Cancer (IARC) Static and extremely low-frequency electric and magnetic fields, Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 80, Non-Ionizing Radiation, Part 1, Lyon : IARC Press, 2002.

[4] Clinard F., Deschamps F. et al., Évaluation de l'exposition aux champs magnétiques dans les habitations situées à proximité des lignes de transport de l'électricité en France. Environnement Risque et Santé (ERS), 2 :111-118, 2004.

[5] http ://www.copublications.greenfacts.org/fr/champs-electromagnetiques/index.htm

[6] Feychting M. and Ahlbom A, Magnetic fields and cancer in children residing near Swedish high- voltage power lines, Am J Epidemiol 138 :467-547, 1993.

[7] Olsen J H., Nielsen A, et al., Residence near high voltage facilities and risk of cancer in children, Bmj, 307 :891-895, 1993.

[8] Verkasalo P K., Pukkala E, et al., Risk of cancer in Finnish children living close to power lines, Bmj, 307 :895-903.

[9] Tynes T. and Haldorsen T, Electromagnetic fields and cancer in children residing near Norwegian high-voltage power lines, Am J Epidemiol, 145 :219-244, 1997.

[10] McBride ML., Gallagher RP., Theriault G., Armstrong BG., Tamaro S., Spinelli JJ. et al., Powerfrequency electric and magnetic fields and risk of childhood leukemia in Canada. Am J Epidemiol, 149 :831-842, 1999.

[11] McCurdy A., Wijnberg L., Loomis D., Savitz D., Nylander-French LA., Exposure to extremely low frequency magnetic fields among working women and homemakers. Ann Occup Hyg, 45 :643-650, 2001.

[12] Linet M., Hatch EE., Kleinerman RA., Robison LL., Kaune WT., Friedman DR., Severson RK., Haines CM., Hartsock CT., Niwa S., Wacholder S., Tarone RE., Residential exposure to magnetic fields and acute lymphoblastic leukemia in children. The New England Journal of Medicine, 337 :1-7, 1997.

[13] UKCCS, Exposure to power-frequency magnetic fields and the risk of childhood cancer. UK Childhood Cancer Study Investigators. Lancet, 354 :1925-1931, 1999.

[14] Ahlbom A, Day N, Feychting M, Roman E, Skinner J, Dockerty J et al., A pooled analysis of magnetic fields and childhood leukaemia. Br J Cancer, 83 :692-698, 2000.

[15] Draper G., Vincent T., Kroll M E., Swanson J., Childhood cancer in relation to distance from high voltage power lines in England and Wales : a case-control study. British Medical Journal, 330 :1279-1359, 2005.

[16] Kabuto M., Nitta H., Yamamoto S., Yamaguchi N., Akiba S., Honda Y. et al. Childhood leukemia and magnetic fields in Japan : a casecontrol study of childhood leukemia and residential power-frequency magnetic fields in Japan. Int J Cancer, 119 :643-650, 2006.

[17] Green LM. et al., Childhood leukemia and personal monitoring of residential exposures to electric and magnetic fields in Ontario, Canada. Cancer Causes and Control, 10 :223-243, 1999.

[18] Schüz J., et al., Extremely low frequency magnetic fields in residences in Germany. Distribution of measurements, comparison of two methods for assessing exposure, and predictors for the occurrence of magnetic fields above background level. Radiat Environ Biophys, 39 :233-240, 2000.

[19] Friedman D.R. et al., Childhood exposure to magnetic fields : residential area measurements compared to personal dosimetry. Epidemiology, 7 :151-155, 1996.

[20] Deadman JE, Armstrong BG, McBride ML, Gallagher R, Thériault G. Exposures of children in Canada to 60-Hz magnetic and electric fields. Scand J Work Environ Health 25 :368-375, 1999.

[21] Gauvin D., Paradis G., Legris M., Levallois P., Niveaux de champ magnétique en milieu scolaire résultant de l'utilisation d'un plancher électrique chauffant. Direction de santé publique de Québec, Institut national de santé publique de Québec, 2003.

[22] International Commission on Non-Ionising Radiation Protection (ICNIRP), Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). Health Phys, 74 :494-522, 1998.

[23] Council of the European Union, Recommendation on the limitation of exposure of the general public to electromagnetic fields (0 Hz to 300 GHz), 1999/519/EC.

[24] European Parliament and Council of the European Union, Directive on the minimum health and safety requirements regarding the exposure of workers to the risks arising from agents (electromagnetic fields), 2004/40/EC.

[25] Institute of Electrical and Electronics Engineers (IEEE). Subcommittee III of Standards Coordinating Committee 28, IEEE Standards Department. IEEE PC95.6-2002 standard for safety levels with respect to human exposure to electromagnetic fields, 0 to 3 kHz. New York : Institute of Electrical and Electronics Engineers, Inc., 2002.

[26] Kheifets L., Sahl JD., Shimkhada R., Repacholi MH., Developing policy in the face of scientific uncertainty : Interpreting 0,3 microT or 0,4 microT cutpoints from EMF epidemiologic studies. Risk Anal, 25 :927-935, 2005.

[27] Kheifets L., van Deventer TE., Lundell G., Swanson J., Le principe de précaution et les champs électriques et magnétiques : mise en oeuvre et évaluation. Environnement, Risques et Santé 5 :43-53, 2006.

[28] Magne,I., Azoulay,A., Lambrozo,J., Souques,M. Comparison of magnetic field meters used for ELF exposure measurement, BEMS, 2006.

[29] Benjamini Y., Hochberg Y., Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57 :125-133, 1995.

[30] Capéraà P., Van Cutsem B., Méthodes et modèles en statistique non paramétrique - Exposé fondamental, Presses de l'Université de Laval et Dunod, 1988.

[31] Aïvasian S., Enukov I., Mechalkine L., Élément de modélisation et traitement primaire des données, Mir, 1986.

[32] Mann H. B., Whitney D. R., On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18 :50-60, 1947.

[33] Hoeffding W., A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematical Statistics, 19 :293-325, 1948.

[34] Bulle T., Comparaison de populations - Tests non paramétriques et analyse de variance, Masson, 1990.

[35] David F., Constructing confidence sets using rank statistics. Journal of the American Statistical Association 67, 687-690, 1972.

[36] Siegel S., Castellan Jr., Nonparametric statistics for the behavioral sciences, McGraw-Hill Inc, 1988.

[37] Howell D., Méthodes statistiques en sciences humaines, De Boeck Université, 1998.

[38] Myers, R. H., Classical and Modern Regression with Applications. Duxbury Press, Belmont, 1990.

[39] Hastie, T. J. et Tibshirani, R. J., Generalized Additive Models. Chapman and Hall, Londres, 1990.

[40] Cleveland, W. S., Robust locally-weighted regression and smoothing scatter-plots. Journal of the American Statistical Association, 74 :829-836, 1979.

[41] Eubank, R. L., Nonparametric Regression and Spline Smoothing. Marcel Dekker, Inc, New York, 1999.

[42] Wood S.N., Generalized Additive Models, An introduction with R, Chapman and Hall/CRC, 2006.

[43] Wood S.N., Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Amer. Statist. Ass., 99 :673-686, 2004.

[44] Wahba G., Bayesian confidence intervals for the cross validation smoothing spline, Journal of the Royal Statistical Society, Series B, 45 :133-150, 1983.

[45] Silverman, B. W., Some Aspects of the Spline Smoothing Approach to Non- Parametric Regression Curve Fitting (with discussion). Journal of the Royal Statistical Society Series B, 47 :1-52, 1985.

[46] Revuz D., Probabilités, Hermann,1997.